Université de Montréal

Limit Order Books in Statistical Arbitrage and Anomaly Detection

par

Cédric Poutré

Département de mathématiques et de statistique

Faculté des arts et des sciences

Thèse présentée en vue de l'obtention du grade de

Philosophiæ Doctor (Ph.D.)

en Mathématiques

Orientation mathématiques appliquées

Avril 2023

Université de Montréal

Faculté des arts et des sciences

Cette thèse intitulée

Limit Order Books in Statistical Arbitrage and Anomaly Detection

présentée par

Cédric Poutré

a été évaluée par un jury composé des personnes suivantes :

Maciej Augustyniak
(président-rapporteur)

Manuel Morales
(directeur de recherche)

Georges Dionne
(codirecteur)

Philippe Gagnon
(membre du jury)

Alexandre F. Roch
(examinateur externe)

Résumé

Cette thèse propose des méthodes exploitant la vaste information contenue dans les carnets d'ordres (LOBs). La première partie de cette thèse découvre des inefficacités dans les LOBs qui sont source d'arbitrage statistique pour les traders haute fréquence. Le chapitre 1 développe de nouvelles relations théoriques entre les actions intercotées afin que leurs prix soient exempts d'arbitrage. Toute déviation de prix est capturée par une stratégie novatrice qui est ensuite évaluée dans un nouvel environnement de *backtesting* permettant l'étude de la latence et de son importance pour les traders haute fréquence. Le chapitre 2 démontre empiriquement l'existence d'arbitrage *lead-lag* à haute fréquence. Les relations dites *lead-lag* ont été bien documentées par le passé, mais aucune étude n'a montré leur véritable potentiel économique. Un modèle économétrique original est proposé pour prédire les rendements de l'actif en retard, ce qu'il réalise de manière précise hors échantillon, conduisant à des opportunités d'arbitrage de courte durée. Dans ces deux chapitres, les inefficacités des LOBs découvertes sont démontrées comme étant rentables, fournissant ainsi une meilleure compréhension des activités des traders haute fréquence. La deuxième partie de cette thèse investigue les séquences anormales dans les LOBs. Le chapitre 3 évalue la performance de méthodes d'apprentissage automatique dans la détection d'ordres frauduleux. En raison de la grande quantité de données, les fraudes sont difficilement détectables et peu de cas sont disponibles pour ajuster les modèles de détection. Un nouveau cadre d'apprentissage profond non supervisé est proposé afin de discerner les comportements anormaux du LOB dans ce contexte ardu. Celui-ci est indépendant de l'actif et peut évoluer avec les marchés, offrant alors de meilleures capacités de détection pour les régulateurs financiers.

Mots clés : Carnet d'ordres; Arbitrage statistique; Trading haute fréquence; Trading algorithmique; Retour à la moyenne; Relation lead-lag; Économétrie financière; Détection d'anomalies dans les séries chronologiques; Apprentissage profond non supervisé; Apprentissage automatique

Abstract

This thesis proposes methods exploiting the vast informational content of limit order books (LOBs). The first part of this thesis discovers LOB inefficiencies that are sources of statistical arbitrage for high-frequency traders. Chapter 1 develops new theoretical relationships between cross-listed stocks, so their prices are arbitrage free. Price deviations are captured by a novel strategy that is then evaluated in a new backtesting environment enabling the study of latency and its importance for high-frequency traders. Chapter 2 empirically demonstrates the existence of lead-lag arbitrage at high-frequency. Lead-lag relationships have been well documented in the past, but no study has shown their true economic potential. An original econometric model is proposed to forecast returns on the lagging asset, and does so accurately out-of-sample, resulting in short-lived arbitrage opportunities. In both chapters, the discovered LOB inefficiencies are shown to be profitable, thus providing a better understanding of high-frequency traders' activities. The second part of this thesis investigates anomalous patterns in LOBs. Chapter 3 studies the performance of machine learning methods in the detection of fraudulent orders. Because of the large amount of LOB data generated daily, trade frauds are challenging to catch, and very few cases are available to fit detection models. A novel unsupervised deep learning–based framework is proposed to discern abnormal LOB behavior in this difficult context. It is asset independent and can evolve alongside markets, providing better fraud detection capabilities to market regulators.

Keywords : Limit order book; Statistical arbitrage; High-frequency trading; Algorithmic trading; Mean-reversion; Lead-lag relationship; Financial econometrics; Time series anomaly detection; Deep unsupervised learning; Machine learning

Contents

List of tables

List of figures

List of acronyms and abbreviations

ADL	*Autoregressive Distributed Lag*
ADR	*American Depositary Receipt*
ADLMLR	*Autoregressive Distributed Lag Multinomial Logistic Regression*
AE	*Autoencoder*
AUPRC	*Area Under the Precision-Recall Curve*
AUROC	*Area Under the Receiver Operating Characterisic Curve*
BATS	*Better Alternative Trading System*
BFGS	*Broyden–Fletcher–Goldfarb–Shanno Algorithm*
CDF	*Cumulative Distribution Function*
CME	*Chicago Mercentile Exchange*
DAX 30	*Deutscher Aktien Index of 30 large companies listed on the Frankfurt Stock Exchange*
ECM	*Error Correction Model*
FTSE	*Financial Times Stock Exchange index of the 100 largest companies listed on the London Stock Exchange*
FX	*Forex*
GAN	*Generative Adversarial Network*
HFT	*High-Frequency Trading*
HFTer	*High-Frequency Trader*

L1	*Level One*
LLR	*Lead-Lag Ratio*
LOB	*Limit Order Book*
LSTM	*Long Short-Term Memory*
MLP	*Multilayer Perceptron*
MIDAS	*U.S. Securities and Exchange Commission's Market Information Data Analytics System*
NYSE	*New York Stock Exchange*
OC-SVM	*One-Class Support Vector Machine*
OLS	*Ordinary Least Squares*
PnL	*Profit and Loss*
RLV	*Realized Local Variation*
SEC	*U.S. Securities and Exchange Commission*
S&P 500	*Standard and Poor's Index of 500 large companies listed in the U.S.*
SPRD	*Relative Spread*
TAQ	*Trades and Quotes*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
TSX	*Toronto Stock Exchange*
TSX 60	*Stock index of TSX's 60 largest companies listed in Canada*
VAR	*Vector Autoregression*

Acknowledgments

Even though my name is solely listed on the title page of this thesis, I would not have been able to achieve this milestone alone. This degree is also the result of the efforts made throughout many years by professors, colleagues, family, and friends from whom I had the privilege to learn and grow as a person.

Firstly, I would like to thank my supervisor, Manuel Morales, who believed in me from the very beginning of my academic journey at the Université de Montréal. His confidence and trust have allowed me to pursue the doctorate and have pushed me to achieve this dream. He has also opened countless doors for me, both professionally and academically, and for that, I am eternally grateful. I am also thankful to my co-supervisor, Georges Dionne, who included me in his research team early in my Ph.D. journey. His great guidance, knowledge, and generosity have made this degree an enjoyable experience. I have also had the pleasure of learning from Gabriel Yergeau, an exceptional mentor. His experience, patience, and constant support have been paramount to the success of this thesis. Thank you.

In addition, I would like to thank the people at IVADO, who make great research projects happen. I also graciously acknowledge the financial support from IVADO, Fin-ML, TMX, Financial Market Surveillance Intelligence Centre, Canada Research Chair in Risk Management, and Mitacs.

Finally, and most importantly, I want to thank my parents and grandparents for their unconditional love, unwavering support, and infinite patience. This thesis could not have happened without you. Another special thank you to my longtime friends (you know who you are) for having the remarkable ability of making me forget about work with your simple presence.

Introduction

Today's financial markets operate at incredibly high speeds, often nearing the speed of light, and human traders have been mostly replaced by computers programmed to make hundreds of decisions in the blink of an eye. In fact, it has been estimated that over 60% of all trading volume in U.S. equity markets was generated by algorithmic trading strategies in 2018, with no end in sight for their dominance.[1] Algorithms and traders alike all interact with central *limit order books* (LOBs) to send orders, trade, and generally gauge other market participants' intentions, in search of profitability.

A LOB is the collection of all outstanding limit orders sent to a trading venue for a single financial asset. It depicts the prices at which participants are willing to buy and sell the security, as well as the volume available at each price level. Whenever a market event changes the LOB's state (e.g., an order is received by the trading venue, an order is canceled or modified), a message from the exchange is sent to every market participant to reflect the new information. The most active financial assets generate millions of messages daily, often in the microsecond range, so the flow of information is humanly impossible to interpret in real time. Even *post factum* analysis is an arduous task. This article–based thesis proposes methods to systematically exploit the vast amount of LOB time series in two research themes: statistical arbitrage, and anomaly detection.

The first two chapters are dedicated to high-frequency trading, in which new statistical arbitrage mechanisms are proposed, and the last chapter focuses on new anomaly detection

[1] "Global Algorithmic Trading Market to Surpass US$ 21,685.53 Million by 2026", *Business Wire*, February 5, 2019, https://www.businesswire.com/news/home/20190205005634/en/Global-Algorithmic-Trading-Market-to-Surpass-US-21685.53-Million-by-2026 (accessed April 17, 2023).

techniques for high-frequency data. The activities of high-frequency traders are still not well understood in practice, so this thesis aims to fill gaps in that literature.

Chapter 1 studies statistical arbitrage in international cross-listed stocks. A new synthetic instrument engineered from cross-listed assets' prices is proposed. The instrument possesses a strong mean-reversion to the exchange rate between the markets where stocks are dual-listed, enabling the derivation of statistical arbitrage bounds. A novel, and practical, high-frequency trading strategy is developed to profit from price deviations outside these bounds. This approach removes the need to quantify the equilibrium level in pairs of assets, while ensuring the mean-reversion of the signal, which are frequent issues in other well-known arbitrage strategies. Furthermore, a methodology accounting for information latency, i.e., the time it takes for a trader to interact with trading venues, is constructed to study its effect on the performance of high-frequency trading strategies, a first in the literature. This chapter demonstrates how acting as a liquidity provider with the proposed strategy can be profitable in North America, whereas using aggressive orders is not, because of market frictions and the high degree of interconnectedness between trading venues in this region.

Chapter 2 examines lead-lag arbitrage at high-frequency. A lead-lag relationship is defined as the phenomenon in which a process is cross-correlated to a second process at a delay. Any lead-lag effect in an asset pair implies the future returns on the lagging asset have the potential to be predicted from past and present prices of the leader. These relationships have been detected in most financial markets, but no clear evidence of their arbitrage potential has been found. This chapter suggests a general framework to test the existence, predictability, and profitability of lead-lag relationships for any pair of assets. It details robust lead-lag estimators applicable to LOB data, known for its asynchronicity and irregular observation times. An econometric model is also developed to exploit lead-lag relationships in level 1 data. Previous methods relied on uniform sampling of the LOB with previous–tick interpolation, whereas the proposed model can employ LOBs as they are observed, generating more accurate return predictions. Because of the significant trading costs associated with market orders, the few existing lead-lag arbitrage strategies are not profitable at high-frequency. To circumvent this problem, Chapter 2 presents a high-frequency trading strategy based on limit orders to capture statistical arbitrage opportunities at minimal cost, using the forecasting ability of the new econometric model. The strategy is empirically able to profit from lead-lag arbitrage in cross-listed stocks, even when important market frictions are considered, thus finally demonstrating their economic value.

Chapter 3 investigates the applicability of machine learning techniques in the detection of anomalies in high-frequency markets. Securities regulators, who constantly monitor trading activities to unveil potential infractions, still perform their investigations manually with the help of deterministic rule–based algorithms. The level of trading activity in modern markets poses a heightened risk of fraudulent orders going unnoticed in the vast quantity of data generated daily. Most of earlier fraud detection methods worked on daily financial data, but switching to LOBs allows to detect the exact fraud time, thus dramatically decreasing the analysis burden of regulators. In this chapter, time series of features are engineered from LOB data to detect potential fraudulent algorithmic patterns. A novel unsupervised deep learning model based on the recent Transformer architecture is devised to capture normal trading behaviors, instead of relying on supervised techniques that would otherwise be unreliable in this context because of the scarcity of real fraud cases. The proposed framework is asset independent and can evolve alongside the market in which it is deployed, hence increasing the detection capabilities of market regulators without necessitating prior knowledge of trade–based manipulation patterns. The method's effectiveness is shown on frauds simulated by a new random manipulation scenario generator.

First Article.

International High-Frequency Arbitrage for Cross-Listed Stocks

by

Cédric Poutré[1], Georges Dionne[2], and Gabriel Yergeau[2]

([1]) Department of Mathematics and Statistics
Université de Montréal
([2]) Department of Finance
HEC Montréal

This article was submitted to International Review of Financial Analysis and was accepted for publication.

The main contributions of Cédric Poutré for this article are presented.

- Creation of the theoretical models and mathematical derivations;
- Design of the trading strategies;
- Production of numerical results;
- Writing most of the manuscript.

Georges Dionne helped with the manuscript. Gabriel Yergeau had the initial general idea behind this article and also helped with the manuscript and data integration.

Abstract. We explore high-frequency arbitrage activities on international cross-listed stocks and develop a methodology to study the effect of information latency in high-frequency trading. We derive statistical arbitrage bounds for a mean-reverting synthetic instrument engineered from cross-listed stock prices, and propose a new strategy that takes advantage of price deviations outside these bounds. Market frictions such as trade costs, inventory control, and arbitrage risks are considered. The strategy is tested with cross-listed stocks involving three exchanges in Canada and the United States in 2019. The annual net profit with the limit order strategy is around US$6 million, whereas the market order version is not profitable because of the great interconnectedness between exchanges in our data.

Keywords: Latency arbitrage; High-frequency trading; Statistical arbitrage; Limit order book; Cross-listed stocks; Supervised machine learning

1. Introduction

We study the profitability of arbitrage activities on international cross-listed stocks in the context of North American markets. Our main research question is as follows: Is international high-speed arbitrage profitable for High-Frequency Traders (HFTers) under strong competition and when all potential arbitrage costs and risks are considered?

Stock exchanges in different countries often use distinct market microstructures, whereas many large public firms employ cross-border listing to reduce their cost of capital and increase their access to liquidity. The current market structure of stock exchanges in North America and Europe is very competitive, fragmented, and fast (Biais and Woolley [7]; Goldstein et al. [34]; Jones [38]; O'Hara [47]; Wah [60]). Changes in regulation, particularly the Regulation NMS in the US and the IIROC rules in Canada,[2] led to an increase in the number of trading venues, thus further fragmenting financial markets (Chao et al. [14]; Garriott et al. [32]). In 2019, there were more than twenty designated exchanges in North America, and the competition related to trading fees, rebates, and colocation fees has increased significantly in recent years (Reuters [50]).

[2] Regulation NMS in the US: SEC Exchange Act Release No. 34-51808 (June 9, 2005). IIROC rules in Canada: CSE Trading Rules and the Universal Market Integrity Rules, of the Investment Industry Regulatory Organization of Canada (IIROC, 2015). See also The MiFID Directive in Europe: Directive 2004/39/EC of the European Parliament and of the Council of April 21, 2004 on markets in financial instruments

The existence of multiple venues means that the price of a given asset need not always be the same across all venues for very short periods of time, opening the door to high-speed arbitrage across markets (Foucault and Biais [26]; O'Hara [47]). Given that this form of arbitrage can be done by creating portfolios that result from spatial arbitrage, traders must appraise intra-market liquidity and analyze the assets' serial correlation. Nonetheless, serial correlation dissipates very quickly, which further increases the possibility of high-speed spatial arbitrage (Budish et al. [12]).

In a market fragmentation context, traders need to search for liquidity across many venues in the same country or across countries. High speed can be crucial when there is strong competition. The ability of HFTers to enter and cancel orders very rapidly makes it hard for many traders to discern where liquidity really exists, which creates more opportunities for HFTers to exploit profitable trading opportunities.

International latency arbitrage opportunities may also arise because of different market models used in local exchanges, different regulations, transient supply and demand shocks, and the arrival of new local information generating asynchronous adjustments in asset prices. These arbitrage possibilities terminate either when an arbitrageur exploits the new opportunity, or when market makers update their quotes to reflect the new information (Foucault et al. [27]). However, local market makers are not always harmonized in real time. High-speed international arbitrage may then benefit all market participants (those with and without high speed) by reducing inter-market bid-ask spreads, a measure of market quality (Hendershott et al. [37]; Riordan and Storkenmaier [51]). As a result, HFTers may even become inter-market makers who provide liquidity with their arbitrage activities, as we demonstrate in this paper.

Whereas arbitrage forces should drive prices to attain an equilibrium, an exchange that acts as a price leader could attract a significant portion of order flow if the adjustment takes time. For example, empirical evidence suggests that prices on Canadian and U.S. exchanges mutually adjust for Canadian–based cross-listed stocks (Chouinard and D'Souza [18]; Eun and Sabherwal [25]).

Considering a cross-country environment, we revisit latency arbitrage strategies, and propose a new model of international mean-reverting arbitrage with FX rate hedging. We are the first to introduce a synthetic instrument, engineered from cross-listed stock prices, that seeks to replicate the exchange rate between currencies. As we will show, this instrument possesses a strong mean-reversion to the actual exchange rate observed in currency futures. Using this property, we derive statistical arbitrage bounds that allow a high-frequency trader to find statistical arbitrage opportunities in cross-listed stocks prices. Taking positions in currency futures also protects the high-frequency trader from currency risk. The earlier literature mainly relied on cross-markets to seek arbitrage opportunities (Budish et al. [12]; Wah [60]), which is only a subset of the opportunities that can be found with our model.

Our strategy is a hybrid between triangular arbitrage (Spraos [56]) and pairs trading (Gatev et al. [33]). Indeed, it relies on the equilibrium of currency instruments (the synthetic instrument and the currency futures), like in triangular arbitrage, but it also trades mainly on a pair of assets (cross-listed stocks) whenever their prices diverge from a historical equilibrium, like in pairs trading. In practice, traders face two problems when considering triangular arbitrage or pairs trading strategies. First, they need to determine the equilibrium-level threshold of the combined positions, which is essential to know *ex ante* the positions' exit point and to determine the expected economic value of any arbitrage opportunity. Second, traders need to ensure that the process resulting from the combined positions will indeed correspond to a mean-reversion process. To the best of our knowledge, we are the first to directly address these two issues simultaneously, and such a strategy has never been proposed.

The present study is also the first to examine stocks' cross-country mean-reverting arbitrage with FX rate hedging. We adopt the perspective of a unique temporal frame of reference, which means that we synchronize the data feeds of exchange venues and explicitly consider the latency that comes from the transmission of information between them and the data processing time. This approach, coupled with the inclusion of trading costs and trading risks in our methodology, generates more realistic results than those obtained in previous studies.

The strategy signals when prices of cross-listed stocks deviate enough from their relative equilibrium that an economically viable arbitrage opportunity occurs. We construct a portfolio of synthetic instruments from pairs of cross-listed stocks of the same company traded on two exchanges and compute their relative spread (SPRD), defined as the ratio of the stock prices (our synthetic future) and a hedging position in the equivalent currency futures. The relative spread deviation resulting from a variation between the synthetic instrument and the hedging instrument is expected to be mean-reverting. We analyze this intraday reverting behavior in detail for each pair of stocks between exchanges. Economically significant deviations of the relative spread from its target value could lead to arbitrage opportunities. We develop different arbitrage strategies to exploit these deviations and to demonstrate the potential profitability of mean-reverting arbitrage opportunities that exist between international exchanges.

According to Foucault and Moinas [28], empirical studies must consider the effect of trading speed on each component of bid-ask spreads separately. These components are adverse selection costs, inventory costs, and order processing costs. We consider adverse selection costs via non-execution risk. Inventory costs are minimized by applying restrictions on the quantities traded and by precluding overnight positions. Order processing costs are considered via infrastructure and trading platform costs, and fees and rebates are also explicitly quantified.

High-frequency trading (HFT) technologies provide speed and information superiority (Biais et al. [6]; Foucault and Moinas [28]), but they introduce various costs such as high technology costs, trading fees and colocation fees (Baron et al. [5]; Shkilko and Sokolov [55]). Potential important arbitrage profits or realized opportunity costs described in the literature are often based on strong (and sometimes unrealistic) assumptions about the functioning of financial markets. The most prevalent ones being latency costs, direct trading fees, rebates on trading fees, and trading platform, colocation and proprietary data feed costs. Moreover, the closing of positions is not always coherent with market reality. Mean-reversion risk, execution risk, and non-execution risk are additional cost components that may affect arbitrage profits. We propose a methodology to introduce all these costs and adjust our algorithms' performance accordingly.

Given that high-frequency trading is very fast and competitive, the risk that the market will move between the time of observing an arbitrage opportunity and the time of the exchange receiving orders sent by a trader's algorithm (i.e., execution risk when using market orders, non-execution risk when using limit orders) is very high. Latency costs for the transmission and the processing of information may matter when exchanges are distant and assets quoted in different currencies are present. Moreover, because gains per trade for high-frequency traders are relatively small given their short holding periods, trading costs and rebates may be significant in the computation of net profits, especially when considering the enormous quantity of trades performed daily by HFTers. The colocation and the proprietary data feed costs are also significant at many exchanges, although they have decreased due to recent competition between exchanges. In this article, we find that, by overlooking these potential costs, the HFT arbitrage profitability presented in the literature has been overestimated (Budish et al. [12]; Dewhurst et al. [20]; Tivnan et al. [58]; Wah [60], among others).

As Chen and Gau [13] assert, the understanding of arbitrage activity in the empirical research is still limited. To our knowledge, we are the first to quantify the importance and the economic value of providing liquidity in the context of arbitrage while considering the limit order book (LOB) queue positions and limit orders instead of market orders exclusively. Our approach is consistent with the revisited HFT market maker definition proposed by O'Hara [47]: "HFT market making differs from traditional market making in that it is often implemented across and within markets, making it akin to statistical arbitrage."[3] Our mean-reverting strategy is a form of statistical arbitrage.

We test the model across three North American exchanges during the first six months of 2019: the New York Stock Exchange (NYSE) and the Chicago Mercantile Exchange

[3] See also Krauss [40]; Rein et al. [49] on statistical arbitrage.

(CME) in the United States, and the Toronto Stock Exchange (TSX) in Canada. We also discuss how the strategy is generalizable to a much larger trading universe without additional restrictions. As Gagnon and Karolyi [31] note, over 3,000 companies had two or more listings in 2008, highlighting the importance of international arbitrage in market equilibrium. Our results report a net annual profit of about C$8 million (US$6 million) for 2019 for this international arbitrage activity, with 36 profitable cross-listed stocks that can be managed by one trader in a large trading firm. These pairs of stocks were selected from 74 potential cross-listed stocks by using a dynamic decision tree model from machine learning. The gross annual profit was about C$19 million, and the main difference between the gross and the net annual profits is explained by latency in the transmission and processing of information, and the non-execution risk of limit orders. Trading fees were consequently not important, yet rebates were significant. We also show that international arbitrage using market orders is not profitable because of transaction fees and the execution risk associated with latency.

The rest of our paper is organized as follows. Section 2 presents the literature on arbitrage trading with LOB data. Emphasis is put on empirical studies that have estimated the profitability of this trading activity in an HFT environment. Section 3 outlines our strategy based on a mean-reverting model of arbitrage that can be executed with market orders or limit orders. We show the main differences between the two approaches with an emphasis on trading cost and rebates. Section 4 presents the methodology used to study the effect of information latency in HFT, and how we consider the multiple arbitrage costs and risks associated with high-frequency arbitrage. Section 5 details the data from TSX, NYSE and CME and how it is managed. It also documents the real latency costs, as well as the trading fees and rebates, the colocation and the proprietary data feed costs at TSX (the trading location used in the application of this paper). Section 6 is dedicated to our empirical results and Section 7 discusses the performance of our arbitrage strategy. Section 8 concludes the paper.

2. Literature review

Two main issues are at the heart of research on HFT: profitability and fairness in trading. Both are interconnected and require appropriate research approaches that are fundamental to understanding the behavior of trading participants and making adequate policy recommendations when necessary. The structure of financial markets has been radically transformed by new technology over the last 25 years. Liquidity and price discovery now arise in a more complex way, often owing to high speed. These changes have affected the market microstructure and the formation of capital in financial markets. They may also have reduced fairness between market participants, warranting new regulatory rules. However, conclusions on the private net benefits of high-frequency trading and its fairness are not always based on solid academic research, according to Chen et al. [15] and O'Hara [47]. Moreover, the debate about the high-frequency trading arms race is still open (Aquilina et al. [2]; Foucault and Moinas [28]).

Academic interest in latency arbitrage is a relatively recent phenomenon, and available studies have investigated it from different angles. The idea that price dislocations exist in fragmented markets is not new. In fact, contributions from the 1990s highlighted the issue in the U.S., even when market fragmentation was not as prevalent as it is today (Blume and Goldstein [8]; Lee et al. [31]; Hasbrouck [35]). More recent studies on that matter include Shkilko et al. [54] and Ding et al. [21]. Soon after, other articles began mentioning the possibility for high-speed traders to exploit these market anomalies. Foucault and Biais [26] and O'Hara [47] both mention that HFTers can capitalize on latency arbitrage opportunities but they conclude that strong empirical evidence is still necessary.

Hasbrouck and Saar [36] are among the first to investigate trading activities within the millisecond environment. Menkveld [44] and Menkveld [45] analyze the behavior of a HFTer who is a market maker. They show that the market maker reduces price variations for the same stock on different exchanges by doing arbitrage activities across trading venues. Budish et al. [12] document the latency arbitrage opportunities between CME and NYSE from 2005 to 2011. They demonstrate that correlation between a pair of related assets breaks down as speed between quotes increases. They show that these breakdowns roughly yield an average of US$75 million a year from a simple latency strategy of arbitraging the spread of one pair of highly correlated assets: the S&P 500 exchange-traded fund (ticker SPY) traded in New York and the S&P 500 E-mini futures contract (ticker ES) traded in Chicago. That pair of instruments had an average of 800 daily arbitrage opportunities during that period, and the authors notice that the arbitrage frequency tracks the overall volatility of the market, with a higher number of opportunities during the financial crisis in 2008, the Flash Crash on May 6, 2010, and the European crisis in summer 2011.

Budish et al. [12] also find that the median ES-SPY arbitrage opportunities duration declines drastically from 97 milliseconds in 2005 to 7 milliseconds in 2011, which is explained by the high-speed arms race led by HFT firms. The median profits per arbitrage opportunity remain relatively constant over time, even though competition clearly reduced the duration of arbitrage opportunities. Budish et al. [12] mention the latency issue, but in a rather incomplete fashion. Their approach does not consider latencies such as the real information transportation cost between the two exchanges nor the information treatment time of a round trip. They may have overestimated the real profits generated by their trading strategy and underestimated the execution risk since they used market orders in their application. In their study, around 85% of latency arbitrage opportunities had a duration of less than 10 milliseconds in 2011. It is possible that this proportion has grown since then, given the technology developments since 2011. This emphasizes the importance of including new latency assumptions for our more recent period of analysis. Finally, as they mention, their strategy only considers bid-ask spread costs, whereas a richer estimate of arbitrage opportunities must also include, at least, exchange fees, and all latency costs. Their study inspired our paper, which seeks to generalize high-frequency arbitrage between pairs of correlated assets and to incorporate practical aspects that are important barriers to the profitability of statistical arbitrage.

Wah [60] examines latency arbitrage opportunities on a larger scale for cross-listed stocks of the S&P 500 in eleven US stock exchanges in 2014. The strategy uses crossed market prices (i.e., when the bid price in an exchange is higher than the ask price in another exchange for the same stock) to locate arbitrage opportunities documented in MIDAS trades and quotes data from the SEC.[4] Considering one infinitely fast arbitrageur operating on these eleven markets, the author estimates that arbitrage opportunity profits were US$3.03 billion in 2014 for the S&P 500 tickers alone. However, round trip information transportation and information treatment time are not considered in the profitability of the strategy, nor are the other trading costs (except for the bid-ask spread cost, due to the use of market orders). Wah [60] influenced the present study by prompting us to reconsider the latency assumptions made in past papers. A better understanding of latency's importance in high-frequency trading is needed, and this is a central aspect of our contribution.

Tivnan et al. [58] and Dewhurst et al. [20] also examine latency arbitrage on cross-listed stocks in the U.S. National Market System, but with MIDAS data from 2016. These two studies consider actionable dislocation segments in their computations, i.e., latency arbitrage opportunities that last longer than the two-way travel time for a fiber-optic cable between exchanges' servers. At this trading speed, the transportation time assumption is especially important, even more so when exchanges are far apart, as in our application. Dewhurst et al. [20] and Tivnan et al. [58] have a more realistic approach when compared with Wah [60] but they do not consider information treatment time, trading costs, and all trading risks.

3. Methodology

3.1. Arbitrage process

We propose an innovative approach for cross-listed stocks arbitrage between two exchanges with differing currencies. In its simplest form, this approach is based on the identification of mean-reverting arbitrage opportunities from a basket of equities traded on their home exchange (noted as Exchange 1), their cross-listed peers at another exchange (noted as Exchange 2), and the currency-futures contract between the two currencies (noted as Currency 1 and Currency 2) for hedging purposes. The strategy also encompasses the simpler

[4] MIDAS is the U.S. Securities and Exchange Commission's Market Information Data Analytics System

case where the two exchanges are using the same currency. That particular application does not require currency hedging, but still relies on the formulations provided in this paper. We will also discuss how the proposed strategy can be generalized to more than two exchanges and two currencies, thus expanding the overall tradable universe.

We first compute a synthetic instrument calculated as the ratio of the stock's simultaneous prices at Exchange 2 and at Exchange 1 (the synthetic, henceforth) obtained from the combination of opposite positions of the same stock being traded on both exchanges. As for internationally cross-listed stocks, the stock prices share two underlying factors: the firm's fundamental value and the exchange rate (Scherrer [52]). Given that we use the same stock in the two exchanges, the idiosyncratic differences are minimal and should not affect the convergence in pairs trading, contrary to what is often observed with different stocks in the literature (Engelberg et al. [24]; Frazzini et al. [29]; Pontiff [48]).

Second, we hedge the synthetic instrument with an opposite position in the currency future. Defining the relative spread (SPRD) as the ratio of the synthetic over the currency future, we must test its stationarity, a *sine qua non* condition for mean-reverting strategies. At equilibrium, SPRD must converge to a value close to 1.0 for each pair in all trading days, with very few exceptions. Spot and futures prices should diverge slightly, only by the basis value, which accounts for maturity differences in the two instruments.

As a distance criterion, we propose a nonparametric threshold rule adjusted for the strategies' net costs in order to uncover economically relevant opportunities. This is an alternative to standard deviation multiples (Gatev et al. [33]; Stübinger [56]). The chosen distance approach is simple and transparent, and allows for large-scale empirical applications (Krauss [40]).

As market makers on either exchange might not be perfectly integrated, we have to consider the differences between the functioning of the microstructures. These sources of divergence may influence limit order books (depth, granularity, imbalance, and bid-ask spread) and marketable orders (trade intensity and potential directional or bouncing behavior).

Data from geographically distant exchanges may be asynchronous. We propose a synchronization procedure to replicate an arbitrageur's information processing lag. We implement a two–regime shift incurred by transport delays of information to and from the exchange servers, and we correct the timestamps for the exchanges' processing time and matching delays. The synchronization is effective at Exchange 1's colocation server.

Our strategy does not hold overnight positions.[5] This prevents hedging overnight gap risk and tying up capital due to end-of-day margin requirements (Menkveld [44]). This also avoids being forced to unwind positions due to margin squeezes (Brunnermeier and Pedersen [11]). We use the exchanges' appropriate trading fees and rebates to evaluate net arbitrage performances, as well as colocation and trading platform expenses. Details on these costs are provided in Section 5.

3.2. Relative spread

Arbitrage opportunities are identified by constructing a relative spread (SPRD) equal to the ratio of the synthetic spread to the hedging instrument (currency futures):

$$\gamma_t \equiv \frac{S_{2,t}/S_{1,t}}{r_t},$$

where γ_t is the observation of process $\{\Gamma_t\}$ at time t. $S_{1,t}$ and $S_{2,t}$ are the cross-listed stock values at Exchange 1 and Exchange 2, and r_t is the exchange rate computed from the currency hedging instrument's value. We define simultaneous prices as prices from a unique time frame of observation that considers the information transportation and treatment time between trading venues, which is known as latency.

We denote

$$\gamma_t^{Short} = \frac{S_{2,t}^{Bid}/S_{1,t}^{Ask}}{r_t^{Ask}} \text{ and } \gamma_t^{Long} = \frac{S_{2,t}^{Ask}/S_{1,t}^{Bid}}{r_t^{Bid}}$$

[5] In our application, we also considered not closing open positions at market close. But, because of the fast mean-reversion time of the signals, and the fact that we stop opening new positions 15 minutes before market close (see Appendix B for more practical considerations), overnight positions were scarce and small in volume. This modification did not significantly modify the strategy's performance, and is not further analyzed.

as the time series of the short and long relative spreads, where the exponents Bid and Ask are the asset prices on the best bid and ask side. We will denote $\left\{\Gamma_t^i\right\},\ i \in \{Short, Long\}$ as the processes with respective observations $\{\gamma_t^i\}$.

3.3. Market order arbitrage strategy

A potential arbitrage opportunity arises when the synthetic instrument is not in equilibrium with the observable exchange rate at time t, that is when:

$$\gamma_t^i \neq \tau^i,\ i \in \{Short, Long\},$$

where τ^i is the equilibrium value expected for the mean-reverting processes. The arbitrage opportunity ends when the equilibrium is restored at time $t > t'$, where t' is defined as:

$$t' \equiv \underset{s>t}{\arg\min}\{s \mid \gamma_s^i = \tau^i\},$$

supposing that $\{\Gamma_t^i\}$ are continuous. The synthetic is potentially overvalued at t when:

$$\gamma_t^{Long} = \frac{S_{2,t}^{Ask}/S_{1,t}^{Bid}}{r_t^{Bid}} > \tau^{Long}.$$

In that case, since $\left\{\Gamma_t^{Long}\right\}$ is assumed to be mean-reverting, the mispricing can be exploited by shorting $1/\tau^{Long}$ shares of Exchange 2 stock, taking a long position of one share in Exchange 1 counterpart (which means that we short the synthetic), and taking a long position in the currency future of the same value as the Exchange 2 stock position in order to hedge our position, all transactions at time t. Then, we must revert the three positions at time t' using market orders to lock the profit per Exchange 1 stock unit, $P_{t'}$, in Currency 1 at time t':

$$P_{t'} = \frac{1}{\tau^{Long} r_{t'}^{Ask}}\left(S_{2,t}^{Bid} - S_{2,t'}^{Ask}\right) + \left(S_{1,t'}^{Bid} - S_{1,t}^{Ask}\right) + \frac{S_{2,t}^{Bid}}{\tau^{Long} r_{t'}^{Ask}}\left(\frac{r_{t'}^{Bid}}{r_t^{Ask}} - 1\right) - c_{t'}^{Long},$$

where $c_{t'}^{Long}$ measures the trading costs in Currency 1 of all the transactions. Considering that the foreign currency market is known for its high liquidity (Campbell and Huang [13]), it is reasonable to assume a narrow bid-ask spread in currency futures, i.e., $r_t^{Ask} \approx r_t^{Bid}$, and obtain the following approximation:[6]

$$P_{t'} \approx \frac{1}{\tau^{Long} r^{Bid}_{t'}} \Big(S^{Bid}_{2,t} - S^{Ask}_{2,t'} \Big) + \Big(S^{Bid}_{1,t'} - S^{Ask}_{1,t} \Big) + \frac{S^{Bid}_{2,t}}{\tau^{Long} r^{Bid}_{t'}} \Big(\frac{r^{Bid}_{t'}}{r^{Ask}_{t}} - 1 \Big) - c^{Long}_{t'}, \qquad (1)$$

where we have substituted $r^{Ask}_{t'}$ for $r^{Bid}_{t'}$. Supposing a perfect hedge, we only buy a fraction of the currency futures of nominal N_{FX} (in Currency 2) that equals the amount invested in Exchange 2 stock at time t. So only a fraction of the constant futures' trading price is paid on this cost-per-share basis. The trading costs paid for opening and closing our positions in Currency 1 at time t', $c^{Long}_{t'}$, are approximated by:

$$c^{Long}_{t'} \approx 2c_1 + 2\frac{c_2}{\tau^{Long} r^{Bid}_{t}} + 2\frac{c_{FX}}{N_{FX}} \cdot \frac{S^{Bid}_{2,t}}{\tau^{Long}},$$

where c_1 and c_2 are the constant per-share trading fees for market orders on Exchange 1 (in Currency 1) and Exchange 2 (in Currency 2) respectively, and c_{FX} is the per-contract trading costs (in Currency 1) with nominal N_{FX}.

When the three instruments return to equilibrium, the definition of t' implies that:

$$\frac{S^{Ask}_{2,t'} / S^{Bid}_{1,t'}}{r^{Bid}_{t'}} = \tau^{Long} \iff \frac{S^{Ask}_{2,t'}}{\tau^{Long} r^{Bid}_{t'}} = S^{Bid}_{1,\tau'}.$$

Using this last equality in equation (1), we get:

$$P_{t'} \approx \frac{S^{Bid}_{2,t}}{\tau^{Long} r^{Ask}_{t}} - S^{Ask}_{1,t} - c^{Long}_{t'},$$

which means that to generate a positive profit at time t', we at least need to have:

$$P_{t'} > 0 \iff \frac{S^{Bid}_{2,t}}{\tau^{Long} r^{Ask}_{t}} - S^{Ask}_{1,t} > c^{Long}_{t'},$$

and we can rewrite this inequality as

$$\gamma^{Long}_{t} \frac{r^{Bid}_{t} S^{Bid}_{1,t}}{S^{Ask}_{2,t}} \cdot \frac{S^{Bid}_{2,t}}{\tau^{Long} r^{Ask}_{t}} - S^{Ask}_{1,t} > c^{Long}_{t'}$$

[6] For example, the average bid-ask spread was around 1.23 bps for the CAD/USD futures, 0.82bps for the EUR/USD futures, and 1.56bps for the JYP/USD futures at CME in 2015–2016, which results in approximations precise up to 10^{-4}. See CME Group (accessed February 22, 2023). The approximation is necessary to eliminate terms observed at t' in the development of nonparametric thresholds.

$$\iff \gamma_t^{Long} > \tau^{Long} \underbrace{\frac{r_t^{Ask}}{r_t^{Bid}} \cdot \frac{S_{2,t}^{Ask}}{S_{2,t}^{Bid}} \cdot \frac{S_{1,t}^{Ask} + c_{t'}^{Long}}{S_{1,t}^{Bid}}}_{>1 \text{ in usual market conditions}} \equiv \kappa_t^{Over}. \tag{2}$$

Equation (2) gives us a dynamic nonparametric upper threshold κ_t^{Over} indicating when a short position in the relative spread (SPRD) is profitable because it is overvalued considering trading costs and bid-ask spreads when only market orders are used. This profitability holds when there is a return to equilibrium to close the positions. The same logic with opposite positions also holds when the synthetic is potentially undervalued, or when:

$$\gamma_t^{Short} = \frac{S_{2,t}^{Bid} / S_{1,t}^{Ask}}{r_t^{Ask}} < \tau^{Short}.$$

This results in a dynamic nonparametric lower threshold at which a long position in the synthetic is profitable considering trading costs and bid-ask spreads when market orders are used:

$$\gamma_t^{Short} < \tau^{Short} \underbrace{\frac{r_t^{Bid}}{r_t^{Ask}} \cdot \frac{S_{2,t}^{Bid}}{S_{2,t}^{Ask}} \cdot \frac{S_{1,t}^{Bid} - c_{t'}^{Short}}{S_{1,t}^{Ask}}}_{<1 \text{ in usual market conditions}} \equiv \kappa_t^{Under}, \tag{3}$$

where $c_{t'}^{Short} \approx 2c_1 + 2\frac{c_2}{\tau^{Short} r_t^{Bid}} + 2\frac{c_{FX}}{N_{FX}} \cdot \frac{S_{1,t}^{Ask}}{\tau^{Short}}$. Since all positions are closed before end of day, we do not include shorting costs in both $c_{t'}^{Long}$ and $c_{t'}^{Short}$. Once again, the profitability of the strategy holds when there is a return to equilibrium to close the long position in SPRD.

From equations (2) and (3), we have a set of two signals, $\{\gamma_t^{Long}\}$ and $\{\gamma_t^{Short}\}$, where $\gamma_t^{Long} > \gamma_t^{Short}\ \forall t, \implies \tau^{Long} > \tau^{Short}$ in usual market conditions (the best bid price is lower than the best ask price in the same LOB) and with their respective dynamic nonparametric thresholds, $\{\kappa_t^{Over}\}$ and $\{\kappa_t^{Under}\}$, where $\kappa^{Over} > \tau^{Long} > \tau^{Short} > \kappa_t^{Under}$, $\forall t$. The arbitrage strategy can be summarized as follows:

- When γ_t^{Long} crosses κ_t^{Over} from below: short $1/\tau^{Long}$ shares of S_2, long S_1 and long the currency future for the same value as the one invested in Exchange 2 stock,
- When γ_t^{Short} crosses κ_t^{Under} from above: long $1/\tau^{Short}$ shares S_2, short S_1 and short the currency future for the same value as the one invested in Exchange 2 stock,

- Close the positions when the equilibrium is restored at t',
- Repatriate the profits generated at Exchange 2 to Exchange 1 whenever they cross N_{FX}.

Hence, the two series $\{\kappa_t^{Under}\}$ and $\{\kappa_t^{Over}\}$ form a bandwidth around $\{\Gamma_t\}$ that needs to be respected between Exchange 1 and Exchange 2 so that no arbitrage opportunities can occur with market orders. Any violation in these conditions, and a high-frequency arbitrageur can potentially profit from the price deviation.

The strategy does not open a new position as long as the previous one has not been closed. Considering that the elapsed time between opening and closing a position, i.e., $t' - t$, can be large, risk management procedures are put in place to minimize the effect of potentially long mean-reversion periods. See Appendix B for further details.

3.4. Limit order arbitrage strategy

We now switch to limit orders, as paying the bid-ask spread on the three instruments can be very costly. The strategy remains the same as with market orders. The main difference is in the profitability equation used to find entry thresholds. The relative spread is potentially overvalued when:

$$\gamma_t^{Short} = \frac{S_{2,t}^{Bid}/S_{1,t}^{Ask}}{r_t^{Ask}} > \tau^{Short}.$$

In that case, we short SPRD at time t and revert the three positions when the equilibrium of $\{\Gamma_t^{Short}\}$ is restored at time t'. This results in a profit in Currency 1 of:

$$P_{t'} = \frac{1}{\tau^{Short} r_{t'}^{Ask}}\left(S_{2,t}^{Ask} - S_{2,t'}^{Bid}\right) + \left(S_{1,t'}^{Ask} - S_{1,t}^{Bid}\right) + \frac{S_{2,t}^{Ask}}{\tau^{Short} r_{t'}^{Ask}}\left(\frac{r_{t'}^{Ask}}{r_t^{Bid}} - 1\right) - \tilde{c}_{t'}^{Short} \quad (4)$$

per Exchange 1 stock, where $\tilde{c}_{t'}^{Short}$ has the same formulation as $c_{t'}^{Short}$, but instead of c_1 and c_2 being the per-share trading fees for market orders, they are now per-share trading fees (or trading rebates) for using limit orders.

Employing the same logic as previously used to obtain the nonparametric entry thresholds κ_t^{Over} and κ_t^{Under}, we find that the dynamic upper threshold indicating a profitable short position in our relative synthetic spread using limit orders is given by:

$$\gamma_t^{Short} > \tau^{Short} \underbrace{\frac{r_t^{Bid}}{r_t^{Ask}} \cdot \frac{S_{2,t}^{Bid}}{S_{2,t}^{Ask}} \cdot \frac{S_{1,t}^{Bid} + \tilde{c}_{t'}^{Short}}{S_{1,t}^{Ask}}}_{\text{multiplicative term}} \equiv \tilde{\kappa}_t^{Over}, \tag{5}$$

and the dynamic lower nonparametric threshold for long positions in our relative synthetic spread using limit orders is given by:

$$\gamma_t^{Long} < \tau^{Long} \underbrace{\frac{r_t^{Ask}}{r_t^{Bid}} \cdot \frac{S_{2,t}^{Ask}}{S_{2,t}^{Bid}} \cdot \frac{S_{1,t}^{Ask} - \tilde{c}_{t'}^{Long}}{S_{1,t}^{Bid}}}_{\text{multiplicative term}} \equiv \tilde{\kappa}_t^{Under}. \tag{6}$$

Notice that the term multiplying the equilibrium level in equation (2) is always greater than the multiplicative term in equation (5). This means that arbitrage opportunities are available at a lower level of γ_t^{Short} with limit orders, and thus should be more frequent. This is true since limit orders greatly reduce the costs related to the strategy. The same observation can be made for the nonparametric thresholds for long positions of equations (3) and (6): Limit orders push the entry thresholds to a more easily attainable level compared with market orders.

From equations (5) and (6), we have a set of two signals, $\{\gamma_t^{Short}\}$ and $\{\gamma_t^{Long}\}$ with their respective dynamic nonparametric thresholds, $\{\tilde{\kappa}_t^{Over}\}$ and $\{\tilde{\kappa}_t^{Under}\}$. The arbitrage strategy can be summarized as follows, supposing the use of fractional shares:

- When γ_t^{Short} crosses $\tilde{\kappa}_t^{Over}$ from below: short $1/\tau^{Short}$ shares of S_2, long S_1 and long the currency future for the same value as the one invested in Exchange 2 stock,
- When γ_t^{Long} crosses $\tilde{\kappa}_t^{Under}$ from above: long $1/\tau^{Long}$ shares S_2, short S_1 and short the currency future for the same value as the one invested in the Exchange 2 stock,
- Close the positions when the equilibrium is restored at t',
- Repatriate the profits generated at the Exchange 2 to the Exchange 1 whenever they cross N_{FX}.

In the strategy's implementation, only round lots are used to reduce trading costs. See Appendix B for further details.

3.5. Strategy at the portfolio level and aggregate hedging

Consider a universe Ω of N cross-listed stocks on Exchange 1 and Exchange 2, $|\Omega| = 2N$. We wish to execute the cross-listed stocks arbitrage strategy defined in the previous sections on every pair contained in that universe. This extension is applicable to both market and limit orders, and is important for the application of the two previous strategies.

Due to the development of our strategy, aggregating every position in a single portfolio offers a built-in hedging effect against movements in the exchange rate whenever positions are open in both $\{\Gamma_t^{Short}\}$ and $\{\Gamma_t^{Long}\}$, because the aggregated position in Exchange 2's market is reduced compared to the sum of the absolute position in every independent pair. The hedge can be optimized with the use of currency futures, and this section explores that extension.

Let us define $\nu_{1,t}^{(n)},\ \nu_{2,t}^{(n)}\ \in \mathbb{R},\ n \in \{1,2,\ldots,N\}$ the size of the position in the cross-listed stock n in both markets at time t. A position is long when the size is positive, short when the size is negative, and the size is zero when no position is open in the asset. Let us also define the total non-repatriated profits, in their respective currency, generated at Exchange 2 and the FX Exchange at time t respectively by $G_{2,t},\ G_{FX,t} \in \mathbb{R}$. Hence, the portfolio's exposures in Currency 1 at Exchange 1, Exchange 2 and FX Exchange at time t are given by:

$$\begin{aligned} V_{1,t} &= \sum_{n=1}^{N} \nu_{1,t}^{(n)} S_{1,t}^{(n)}, \\ V_{2,t} &= \sum_{n=1}^{N} \nu_{2,t}^{(n)} \frac{S_{2,t}^{(n)}}{r_t} + \frac{G_{2,t}}{r_t}, \\ V_{FX,t} &= \frac{\nu_{FX,t}^{*} N_{FX}}{r_t} + G_{FX,t}, \end{aligned}$$

where $\nu_{FX,t}^{*} \in \mathbb{R}$ is the optimal position size in the currency futures at time t that we are trying to obtain. The total value of the portfolio in Currency 1, V_t, is given by:

$$V_t = V_{1,t} + V_{2,t} + V_{FX,t}.$$

By taking a position in the currency futures that is the inverse of the position in Exchange 2, we get:

$$V_{FX,t} = -V_{2,t} \iff \nu^*_{FX,t} = -r_t \frac{V_{2,t} + G_{FX,t}}{N_{FX}}, \tag{7}$$

which results in a neutral aggregated position in Exchange 2's market: $V_{2,t} + V_{FX,t} = 0$. The portfolio's value is now simply given by $V_t = V_{1,t} \implies \frac{\partial V_t}{\partial r_t} = \frac{\partial V_{1,t}}{\partial r_t} = 0$, assuming the mathematical independence between Exchange 1 stocks' prices and the exchange rate. In the universe Ω, a portfolio invested in cross-listed stock pairs following the proposed strategy for every pair achieves an optimal hedge against currency risk at any time t when that portfolio has a neutral aggregated position in Currency 2. If the aggregated position in Exchange 2 stocks is not neutral, a position of $\nu^*_{FX,t}$ contracts can be taken in the currency futures to get a perfect hedge.

The hedging of the portfolio is done by rebalancing the position in the currency futures to the optimal value, if necessary, whenever positions are open or closed in pairs of cross-listed stocks, compared with the pairwise strategy which requires taking the inverse of the position taken at Exchange 2 at every arbitrage opportunity.

3.6. Generalization of the strategy beyond two stock exchanges and a single exchange rate

The proposed strategy and the formulated arbitrage signals can be applied to more general trading environments. Indeed, the arbitrage signals $\{\Gamma_t\}$ defined in this section can be computed for any cross-listed asset pair between any two exchanges and any currency for both assets (shared or not) without any modification. The global tradable universe for which the proposed strategy can be applied to is thus quite large, as discussed in the introduction. Different additional trading environments where the strategy can be applied are presented.

The first additional trading environment would be when there are two exchanges and a single currency for the cross-listed asset's pair. This can be done by setting $r_t = r_t^{Bid} = r_t^{Ask} = 1$, $\forall t$ and ignoring the currency hedging instrument. The signals are thus solely based on the equilibrium between the two microstructures, which corresponds to the model of Budish et al. [12]: Whenever a sudden jump occurs in one of the two stocks, the correlation between them breaks down and an arbitrage opportunity potentially opens up. The arbitrage signals proposed in this article consider both the closing conditions and the trading costs associated with sending orders to seize the arbitrage opportunity.

The second trading environment would be when there are more than two exchanges and a single currency for the cross-listed assets. Once again, this can be done by using the same constraint on r_t and ignoring currency hedging as previously discussed. But a second constraint needs to be put in place to select which arbitrage opportunity to capture, whenever multiple opportunities occur at the same time for the same asset and exchange. This is necessary since each asset can be part of more than two exchanges, so multiple cross-listed pairs can contain the given asset. In that case, only the cross-listed pair with signal $\{\Gamma_t\}$ that is the furthest from equilibrium τ is executed (i.e., the pair with the maximum expected profitability). This relates closely to the model of Wah [60], but the author did not consider latency, inventory management, nor any trading cost.

The final case would be when there are more than two exchanges and multiple exchange rates hosted by any number of forex exchanges. The trading signals $\{\Gamma_t\}$ can be computed for every combination of cross-listed asset pair and their applicable exchange rate. As in the previous case, multiple arbitrage opportunities can happen at the same time for the same asset at a single exchange. Again, only the pair with signal $\{\Gamma_t\}$ that is the furthest from its equilibrium τ is executed for that particular asset. To the best of our knowledge, this has not been studied in the literature yet.

Overall, by adding simple constraints to the proposed strategy, either on the observable exchange rate r_t, currency hedging, or on the selection of arbitrage opportunities computed by our signals $\{\Gamma_t\}$, the strategy can be applied to any asset pair.

4. Latencies, arbitrage costs, and arbitrage risks

4.1. Latencies and arbitrage costs

A factor of interest in this contribution is latency. In trading terms, latency refers to the time it takes for an agent to interact with the market. We closely follow the measure of latency proposed by Hasbrouck and Saar [36], which is based on three components: the time it takes for a trader to learn about an event, generate a response, and have the exchange act on that response (see also Foucault and Moinas [28]). We split that definition into two separate quantities so that we can have more granularity on the impact of latency on the high-frequency trading strategies.

The first quantity of importance is the latency of a message from any exchange to Exchange 1, which includes the one-way transportation time of the information to Exchange 1, and the information treatment time needed by the agent's servers colocated at that last exchange. The second quantity of importance is the latency of a message from Exchange 1 to another exchange, which is comprised of the one-way transportation time of information from Exchange 1 to the receiving exchange, and the matching engine delay of that last exchange.

Information treatment time refers to the timespan required to receive and analyze incoming information from the exchanges, followed by the decision to trade or not. Exchanges server procedure considers information reception at the exchange gates, LOB positioning or matching of an incoming limit order (with the LOB) and issuing traders' confirmation to the server gates. Round-trip latency measures the total latency delay for a message between two exchanges.

A two-regime model associated with regular and extreme market conditions based on quote and trade message intensity is applied. Regime shifts, from the regular state to the extreme one, are often due to bursts in the events stream, phenomena well documented in the literature (Dixon et al. [22]; Friederich and Payne [30]; Menkveld [45]; Shkilko and Sokolov [55]). To help in recreating this behavior, a latency regime variable depending on the number of messages a certain exchange received in the last millisecond on a per–asset basis is used. This quantity is a good proxy of an exchange's server traffic, which has a positive relationship with computational delays occurring during the information treatment time and the matching engine time components of latency. The regular regime generates a minimal, baseline value of the latency that exists between two exchanges and a bonus on that minimal latency is added for the extreme regime.

The latency regime variable for a given asset remains in its regular state up to a certain static threshold for the number of messages in a single millisecond for that asset, which is set as the 95^{th} percentile of its empirical distribution. Let's define $q^i_{95\%}$ as the 95^{th} percentile of the empirical distribution of the number of messages in one millisecond for asset i, and define q^i_j the number of messages during the millisecond preceding and ending at message $j \in [1,N^i]$ where N^i is the total number of messages for asset i during the full period. Let's also define $L^i_j \in \{\text{Regular, Extreme}\}$ the latency regime of asset i at message j. Then, its value is computed as follows:

$$L^i_j = \begin{cases} \text{Regular,} & \text{if } q^i_j < q^i_{95\%} \\ \text{Extreme,} & \text{if } q^i_j \geq q^i_{95\%} \end{cases} \quad \forall i,j.$$

By adding the corresponding latency to the original exchange timestamp of every message, the data feeds of geographically distant exchanges can be approximately synchronized into a single point of observation (e.g., Exchange 1) as they would be in practice because of the natural and technological limits of information propagation. The methodology emulates that relativistic effect so that what is observed by the trading algorithm at any point is a past state of markets. The same idea applies when the algorithm sends an order to a given exchange. The corresponding latency is added so that the agent does not interact immediately with that exchange. This makes it possible to study the influence of latency on the performance of high-frequency trading strategies.

4.2. Arbitrage risks

4.2.1. Execution risk.

The choice between limit and market orders relies, in part, on the difference between non-execution risk and execution risk (Brolley [10]; Dugast [23]; Kozhan and Tham [39]; Liu [42]; Mavroudis [43]). To empirically solve this trade-off, we first evaluate our algorithm's performance using market orders exclusively. As we will see, using only market orders leads to a negative economic value with our data in the sense that the cost of immediacy (conceding the bid-ask spread) cannot be borne by the arbitrageur in the vast majority of trades. This high cost also results in a very low number of potential arbitrage opportunities, since the divergence of SPRD is rarely large enough to compensate it. We then constrain our algorithm to limit orders, except for the liquidation of positions to avoid overnight exposures. We also use marketable limit orders (i.e., liquidity taking limit orders at the first level) to offset unexecuted legs. There remain two additional risks.

4.2.2. Non-execution risk.

We evaluate non-execution risk costs by managing the LOB queuing priorities. We mitigate the risk of non-execution by dynamically keeping our limit orders to the LOB's level one. This is implemented conditional on the persistence of an expected profitable arbitrage. Otherwise, we liquidate positions, if any, by issuing marketable limit orders (Dahlström and Nordén [19]).

4.2.3. Mean-reversion risk.

Mean-reversion risk arises after initial positions are taken. It materializes when the circuit breaker timer is triggered (see Appendix B for details). All arbitrage legs are then liquidated via marketable limit orders. As we will see, this risk is very low in our data since the processes $\{\Gamma_t^{Short}\}$ and $\{\Gamma_t^{Long}\}$ are stationary for almost all stocks and trading days.

5. Data, synchronization, TAQ emulator and trading costs

5.1. Data

We use LOB level one data and trade data that we obtained from: the TAQ NYSE OpenBook and the TAQ NYSE Trades historical data timestamped to the microsecond, the CME Market Depth FIX Canadian Dollar Futures historical data timestamped to the nanosecond, and Trades and Quotes Daily historical data from TMX Group timestamped to the nanosecond. All the data sets were timestamped at their respective exchange, and span from January 7, 2019 to June 28, 2019, inclusively. We only select dates where the three exchanges were open, meaning that we remove every holiday from our sample.[7] The timestamps are truncated and rounded to the nearest millisecond above so that potential microscopic errors in the timestamps do not affect the results.

Overall, there are 120 trading days in our data set. We have access to 74 pairs of cross-listed stocks that were listed on both the TSX and the NYSE during at least two weeks of that period. Pairs where one of the stocks got delisted from an exchange at any point are kept in the sample, but the strategy is only applied to periods where both stocks of the pair were listed and active. All cross-listed S&P/TSX 60 stocks are present in our sample during the six months. Table 11 of Appendix A describes every available pair and Table 12 includes their aggregated statistics during the period of analysis.

We use the quarterly CAD/USD futures listed on CME: 6CH9 expiring March 19, 2019; 6CM9 expiring June 18, 2019; and 6CU9 expiring September 17, 2019. We do not use monthly futures because of their smaller open interest. A continuous futures contract is created by concatenating the three futures' data and by adjusting the LOB level one and trade prices of the consecutive contracts so that no jumps are artificially created. The concatenation dates are determined based on the daily transaction volume of consecutive futures. That

[7] TSX: February 18: Family Day; April 19: Good Friday; May 20: Patriot's Day. NYSE and CME: January 21: Martin Luther King Jr. Day; February 18: President's Day; April 19: Good Friday; May 27: Memorial Day.

is, whenever the futures contract with the furthest expiration date generates a significantly higher daily transaction volume than its predecessor and remains more actively traded, we switch to those futures' trades and quotes for the continuous futures used in the strategy. In order to have better hedging capabilities, we employ the Micro CAD/USD futures contract with a nominal of C$10,000, which we approximate by dividing the continuous futures' prices by 10, because of its nominal of C$100,000.

5.2. Data synchronization

The strategy is executed each week, from Monday to Friday, starting at 9:30 a.m. and ending at 4:00 p.m. Eastern Time when the three exchanges are all open to continuous trading. Both TSX and NYSE are in the Eastern Time zone, but CME is in the Central Time zone, one hour behind. Hence, we add an hour to the timestamps of CME data to synchronize the three exchanges' clocks.

5.3. TAQ emulator

The methodology and the different trading strategies are implemented in Deltix QuantOffice, a trading software suite used by multiple traders, which brings them closer to real trading practice. The Deltix trading suite allows to replay the synchronized events of the three stock markets (level one LOB and trades) as they were obtained in streaming by traders. By handling these events and following orders position in the queues, the real-time performance of the strategies can be computed as realistically as possible. This implementation allows trading fees and rebates, latency, and other trading risks and costs all presented above. It confirms the order status (creation, cancelation, or execution) just as it would have happened in streaming trading while considering market frictions and ever-changing market states. Standard reports, such as a trades and performance reports, are generated at the end of a strategy's execution. These are used to compute our results.

Moreover, the individual and aggregated positions can be managed, and the respective Profit and Loss Reports (PnL) can be calculated altogether with performance statistics.

These PnLs represent the economic value of the arbitrage opportunities. Using their performance as a benchmark, the economic impact of latency risk can be evaluated by varying the aforementioned latency parameters. The general rules of the trading and quoting emulator on L1 data, and information on how executions and non-executions occur are detailed in Appendix C.

5.4. Empirical latencies and other costs

Table 1 documents the 2019 latency costs, trading costs, rebates, colocation costs, and proprietary data feed (including the trading platform) costs used in this study. Orders and positions are managed at TSX's colocation premises in Toronto (TSX [59]). Information comes from TSX, NYSE, and CME. Asynchronicities are addressed by adjusting the TSX timestamps based on round-trip transportation time, arbitrageur information processing delays, and exchanges matching engine delays presented in Table 2. Table 1 also documents the fees for the liquidity–removing trades and rebates for the liquidity–providing trades. Colocation costs in Toronto are considered in the monthly portfolio performance estimations, as well as proprietary data feeds. Colocation enables some trading firms to receive updates from the exchange faster than other traders who do not pay for this service.

For both latency regimes, the latency to and from TSX is set as the sum of its components' interval center found in Table 2, for the respective market condition. Latencies are rounded up to the closest integer. Table 3 details the empirical latencies used. Following the methodology introduced in Section 4, the estimation of the empirical distribution of messages per millisecond used a random sample of six weeks, where each sampled week came from a different month contained in our data.

6. Empirical results

The empirical results are presented in two steps. First, the trading strategy performance of Budish et al. [12] applied to our data is computed.[8] The goal of this exercise is to isolate

[8] See Appendix D for the analysis of Wah [60].

Table 1. Arbitrage costs

Definition	Description	Measurement	In Deltix
Information transportation time between exchanges	Transportation time details: Toronto – Chicago: Fiber paths Toronto – New York: Microwave path (regular) Fiber path (extreme situations)	See Table 2	Adjusted raw dataset timestamp fed to Deltix
Information treatment time	Timespan required to receive and analyze incoming information from the exchanges, followed by the decision to trade or not	See Table 2	Adjusted raw dataset timestamp fed to Deltix
Exchange trading fees	TSX member trading fees per share[1]	Removing: \$0.0015 Providing: (\$0.0011)	Applied to matched orders
	NYSE Type A stocks per share[2]	Removing: \$0.00275 Providing: (\$0.00120)	
	CME Globex C/US FX futures per contract[3]	\$100k notional value: \$0.32 \$10k notional value (e-micro): \$0.04	
Colocation cost	Colocation with exchange connectivity rates	Half cabinet (21U, 3 kw maximum): \$5,250 monthly Initial set-up fee: \$5,250 one-time	Included in monthly portfolios performance
Proprietary data feed	TSX & Venture level 1 Distribution Trading use case license	\$4,000 monthly	Included in monthly portfolios performance

[1] https://www.tsx.com/resource/en/1756/tsx-trading-fee-schedule-effective-june-4-2018-en.pdf
[2] https://www.nyse.com/markets/nyse/trading-info/fees
[3] https://www.cmegroup.com/company/clearing-fees.html

the importance of considering latencies, execution risk, and trading costs when evaluating the benefits of HFT arbitrage. It also serves as a benchmark to compare our trading strategies and test the profitability of previously proposed arbitrage on more recent data.

Second, the results from our strategies are presented. It is shown that arbitrage with market orders is not profitable, while arbitrage with limit orders provides net profits when latencies, rebates, exchange fees, and non-execution risk are considered. Other conclusions are discussed.

Table 2. Latencies[1]

Market condition	Exchanges from–to	Transportation time	Information treatment time	Exchanges from–to	Transportation time	Exchange server	Round-trip latency
Regular	TSX–TSX	5 μs	10–70 μs	TSX–TSX	5 μs	100–300 μs	120–380 μs
	NYSE–TSX	2.4 ms	10–70 μs	TSX–NYSE	2.4 ms	100–300 μs	4.91–5.17 ms
	CME–TSX	5 ms	10–70 μs	TSX–CME	5 ms	1–5 ms	11.01–15.07 ms
Extreme	TSX—TSX	5–10 μs	200–500 μs	TSX–TSX	5–10 μs	5–10 ms	5.21–10.52 ms
	NYSE–TSX	4.8–9.6 ms	200–500 μs	TSX–NYSE	4.8–9.6 ms	5–10 ms	14.80–29.7 ms
	CME–TSX	5–10 ms	200–500 μs	TSX–CME	5–10 ms	50–100 ms	60.20–120.50 ms

[1] Latencies are obtained following discussions with a major Canadian financial institution trading actively in Canada and in the United-States. ms: millisecond; μs: microseconds.

Table 3. Latencies used when testing the strategies, depending on the latency regime, the origin of the message and the exchange where the message is sent.

Latency regime	Exchange from–to	Latency	Exchanges from–to	Latency
Regular	TSX–TSX	1 ms	TSX–TSX	1 ms
	NYSE–TSX	3 ms	TSX–NYSE	3 ms
	CME–TSX	6 ms	TSX–CME	8 ms
Extreme	TSX–TSX	1 ms	TSX–TSX	8 ms
	NYSE–TSX	8 ms	TSX–NYSE	15 ms
	CME–TSX	8 ms	TSX–CME	83 ms

6.1. Budish, Crampton and Shim's strategy

This contribution examines arbitrage opportunities between the two largest financial instruments tracking the S&P 500 index: the SPDR S&P 500 exchange-traded fund (ticker SPY) and the S&P 500 E-mini futures contract (ticker ES), using millisecond-level direct feed data from different stock exchanges and CME. The application is consequently very different from arbitrage trading of the same stock in two different exchanges, but some comparisons with our research are important, given that this article suggests strong modifications to the functioning of continuous HFT. The authors first demonstrate that the high correlation between the two securities observed from the bid-ask midpoints breaks down at very high-frequency. This correlation breakdown creates technical arbitrage opportunities estimated at approximately US$75 million of gross profit per year for the two securities alone on all markets where the SPY is traded (not only at the NYSE). Their period of analysis includes many high volatility periods such as the 2007-2009 financial crisis. For a more regular year like 2005, the total gross profit is US$35 million.[9] Verifying from Bloomberg that the share of the NYSE for this market is 25%, the annual gross profit for 2005 is US$8.75 million for the NYSE alone. These numbers represent gross profits because trading fees are not considered, nor are latencies and exchange fees. Only bid-ask spread costs are computed.

The above numbers come from the following market environment: There is no arbitrageur entry in the market over the period considered and a single trader observes variations of the stock price with zero-time delay. There is also zero latency in sending orders and receiving updates from the exchanges. This is a pure continuous trading environment without asymmetric information and inventory costs, where open positions at an exchange can be immediately closed at another exchange with a different asset.

The strategy of Budish et al. [12] is first implemented with their theoretical settings and minor modifications to adapt it to our data. In that sense, prices at NYSE are continuously transferred to CAD following the CAD/USD futures observed at CME. In addition, we use two hypotheses employed in their model: There is an absence of latency and open positions at an exchange can be immediately closed at another exchange, resulting in a trade. Table 4 presents the results obtained on our data with the arbitrage strategy presented in Online appendix A.2 of their article with market orders only. The second column of Table 4 presents the results that are obtained following as closely as possible their theoretical framework. In the next two columns, latency is considered.

We observe, in column 2 of Table 4, that gross profit is limited to C$1.4 million for six months of continuous trading, or about C$2.8 million for a year, which is below the C$10.60 million (US$8.75 million) for the low volatility year of 2005 with their data. Many factors

[9] The CBOE Volatility Index (VIX) of the average closing price was equal to 12.81 in 2005, 32.69 in 2008, and 15.39 in 2019.

Table 4. Budish et al. [12] model on our data

1	2	3	4
Model	**Budish Original**	**Budish Original**	**Budish Original**
		w/ 1x Latency	**w/ 3x Latency**
Latency multiplier	0	1	3
Pair selection	No	No	No
Gross profit	$1,421,685.23	$998,328.25	$1,116,673.07
Loss	$0.00	-$11,492.18	-$18,696.78
Trading fees	-$75,167.39	-$57,973.82	-$67,232.10
Trading rebates	$0.00	$0.00	$0.00
Total net profit	$1,346,517.84	$928,862.25	$1,030,744.19
Mean daily net profit	$11,811.56	$8,147.91	$9,041.62
Median daily net profit	$1,968.76	$1,189.76	$1,219.35
Mean daily net profit per pair, per day	$110.95	$76.54	$84.93
$p - value$ Kolmogorov-Smirnov test[1]		1.00	0.65
1st most profitable day	2019/01/28	2019/01/28	2019/01/28
(date - profit)	$184,196.22	$121,108.28	$127,578.22
5th most profitable day	2019/01/30	2019/01/24	2019/01/24
(date - profit)	$66,060.79	$47,904.13	$50,816.97
1st most unprofitable day	2019/06/24	2019/06/24	2019/06/03
(date - profit)	-$161.55	-$450.32	-$2,222.67
5th most unprofitable day	2019/05/31	2019/06/27	2019/06/24
(date - profit)	-$77.85	-$340.18	-$681.72
Average time in trade[2]	00:00.0	00:00.0	00:00.0
# Net profitable trades	31,762	23,313	29,226
# Net unprofitable trades	1,176	1,336	1,817
# Trades	32,938	24,649	31,043
% Net profitable trades	96.43%	94.58%	94.15%
Average volume per trade	345.63	352.77	326.16
Average net profit per trade	$40.88	$37.68	$33.20
Average profit per net profitable trades	$42.75	$40.75	$36.32
Average profit per net unprofitable trades	-$7.97	-$15.91	-$16.99

[1] $H_0: \ F(x) \leq G(x), \ H_1: \ F(x) > G(x)$. $F(x), \ G(x)$ = cumulative distribution functions of daily net profits from sample 1 and sample 2, respectively: p-value of 1.00 for no latency v.s. 1x latency and 0.65 for 1x latency v.s. 3x latency.

[2] HH:MM:SS.U – hours:minutes:seconds:fractions of a second.

can explain the difference. The main difference is mostly related to the average daily trading activity of the assets. They document 800 daily arbitrage opportunities in their data, while in our data we have 200 daily arbitrage opportunities with their strategy for the 74 stocks.

Introducing trading fees does not significantly affect the profitability in the second column, but some opportunities do not cover the trading costs. The main difference in profitability is obtained when latency is introduced. This effect is observed in the next two columns where the total net profitability drops by more than 30%. The daily net profitability is statistically greater when latency is ignored (see $p - values$). This is explained by the fact that true cross-markets occasions observed at a single geographical point last a shorter amount of time, and some opportunities are now nonexistent compared to a latency–free environment, thus decreasing the number of trades by around 25%. Captured arbitrage opportunities are also less profitable. Comparing the net profitability of column 2 with that of column 3, it can observed that profits were indeed inflated in column 2 because of a simplified market environment.

Another hypothesis was made in the strategy of Budish et al. [12]: Exact opposite positions in different exchanges count as a trade and result in a null inventory in both accounts. The next table, Table 5, does not use this simplified environment, meaning that positions can only be closed with an opposite position at the same exchange with the same stock. The second column does not include latency. The next two columns do.

Table 5. Budish et al. [12] model on our data with practical hypotheses

1	**2**	**3**	**4**
Model	**Budish Practical w/o Latency**	**Budish Practical w/ 1x Latency**	**Budish Practical w/ 3x Latency**
Latency multiplier	0	1	3
Pair selection	No	No	No
Gross profit	$779,282.29	$441,466.25	$666,886.91
Loss	-$789,845.78	-$456,295.76	-$695,876.53
Trading fees	-$11,686.80	-$6,957.61	-$11,089.11
Trading rebates	$0.00	$0.00	$0.00
Total net profit	-$22,250.29	-$21,787.12	-$40,078.73
Mean daily net profit	-$195.18	-$191.12	-$351.57
Median daily net profit	-$5.11	-$44.00	-$49.72
Mean daily net profit per pair, per day	-$1.83	-$1.80	-$3.30
$p - value$ Kolmogorov-Smirnov test[1]		0.18	0.80
1st most profitable day (date - profit)	2019/06/27 $2,473.72	2019/06/28 $2,043.65	2019/06/28 $2,158.73
5th most profitable day (date - profit)	2019/06/20 $1,219.50	2019/06/21 $292.03	2019/06/20 $233.39
1st most unprofitable day (date - profit)	2019/05/15 -$9,698.68	2019/05/15 -$5,221.17	2019/06/03 -$7,570.36
5th most unprofitable day (date - profit)	2019/06/03 -$1,718.65	2019/05/21 -$1,294.97	2019/06/05 -$2,132.83
Average time in trade[2]	126.06:12:08	127.12:57:37	127.14:15:11
# Net profitable trades	974	702	961
# Net unprofitable trades	958	708	987
# Trades	1,932	1,410	1,948
% Net profitable trades	50.41%	49.79%	49.33%
Average volume per trade	585.56	477.63	551.30
Average net profit per trade	-$11.52	-$15.45	-$20.57
Average profit per net profitable trades	$796.17	$625.62	$690.11
Average profit per net unprofitable trades	-$832.69	-$651.09	-$712.53
Total Short Inventory Remaining	$354,467,602.46	$276,237,299.21	$309,494,680.19
Total Long Inventory Remaining	$271,097,081.28	$211,074,656.88	$236,477,971.72

[1] $H_0: \ F(x) \leq G(x), \ H_1: \ F(x) > G(x)$. $F(x), \ G(x) =$ cumulative distribution functions of daily net profits from sample 1 and sample 2, respectively: p-value of 0.18 for no latency v.s. 1x latency and 0.80 for 1x latency v.s. 3x latency.

[2] HH:MM:SS.U – hours:minutes:seconds:fractions of a second.

As can be seen in Table 5, the strategy does not generate any net profits when the hypothesis of a trade occurring when exact opposite positions are taken in two different exchanges is abandoned. The net profitability is even more statistically reduced when latency is considered. Column 3 of Table 5 would be the closest results obtained by a HFT firm using the strategy during our data period. Another salient point is the very large accumulated inventory that needs to be managed. This is attributable to the fact that price discovery primarily occurs on the Canadian exchange (Chouinard and D'Souza [18]; Eun and Sabherwal [25]). Coupled with a positive directional market like in our period, the jumps in prices occurred on the bid side of the book for the Canadian stock first most of the time.[10] This resulted in taking the same short TSX positions and long NYSE positions repeatedly, thus rarely closing previous positions to generate a trade. This shows the importance of inventory management and currency hedging in an international arbitrage context. Overall, by not considering practical trading aspects such as latency or real market functioning, Budish et al. [12] inflated latency arbitrage profitability.

6.2. Market order–based strategy

Using the Augmented Dickey-Fuller test for stationarity, we obtain that both $\{\gamma_t^i\},\ i \in \{Short, Long\}$ time series from January 7, 2019 to June 28, 2019 are stationary for almost all stocks in all trading days where the three exchanges are open at the same time, at a $p-value$ of 1%. Details are presented in Table 13 of Appendix A. Given that these time series are stationary and exhibit strong mean-reversion, τ^i can be defined as the equilibrium level of the mean-reverting processes $\left\{\Gamma_t^i\right\}$.

The main results from the market order strategy are presented in Table 6. This strategy is not profitable because it is too expensive to obtain enough liquidity and orders are subject to execution risk (Loss row). Thus, following our theoretical strategy with market orders is hazardous, especially when latency is considered. Indeed, we also observe, in columns three and four, that increasing latency reduces the net profitability even more and this effect is largely significant in both columns (significant $p-values$). The limited number of trades reflects that TSX and NYSE are very well integrated, because the signals $\{\gamma_t^{Short}\}$ and $\{\gamma_t^{Long}\}$ rarely cross their respective thresholds $\{\kappa_t^{Under}\}$ and $\{\kappa_t^{Over}\}$, resulting in a low amount of potential arbitrage opportunities for a HFTer.

[10] The only exception is TRQ, which dropped by 25% in our period, exhibiting an exact opposite trading behavior.

Table 6. Results with our market orders–based strategy

1	**2**	**3**	**4**
Model	**Market Orders w/o Latency**	**Market Orders w/ 1x Latency**	**Market Orders w/ 3x Latency**
Latency multiplier	0	1	3
Pair selection	No	No	No
Gross profit	\$38,660.35	\$41,508.69	\$41,620.24
Loss	-\$58,361.15	-\$96,751.29	-\$128,442.17
Trading fees	-\$17,890.26	-\$22,121.43	-\$31,985.04
Trading rebates	\$0.00	\$0.00	\$0.00
Total net profit	-\$37,591.06	-\$77,364.03	-\$118,806.97
Mean daily net profit	-\$329.75	-\$678.63	-\$1,042.17
Median daily net profit	-\$18,24	-\$207.53	-\$595.92
Mean daily net profit per pair, per day	-\$4.46	-\$9.17	-\$14.08
$p - value$ Kolmogorov-Smirnov test[1]		1.00	1.00
1st most profitable day (date - profit)	2019/03/06 \$354.30	2019/05/31 \$21.63	2019/05/31 \$51.54
5th most profitable day (date - profit)	2019/06/21 \$196.92	2019/06/17 -\$2.54	2019/04/29 -\$94.49
1st most unprofitable day (date - profit)	2019/01/30 -\$4,053.94	2019/01/16 -\$4,682.15	2019/05/16 -\$4,692.79
5th most unprofitable day (date - profit)	2019/03/26 -\$2,095.20	2019/01/29 -\$3,504.39	2019/01/28 -\$3,785.02
Average time in trade (excl. futures contracts)	00:06:34.41	00:06:37.83	00:04:42.15
Average time in trade[2] (incl. futures contracts)	02:12:28.60	00:59:30.36	00:57:59.53
# Net profitable trades	1,284	1,092	1,590
# Net unprofitable trades	2,130	2,927	4,814
# Trades	3,414	4,019	6,404
% Net profitable trades	37,61%	27.17%	24.83%
Average volume per trade	1,529.78	1,592.15	1,449.57
Average net profit per trade	-\$11.01	-\$19.25	-\$18.55
Average profit per net profitable trades	\$26.46	\$32.92	\$21.99
Average profit per net unprofitable trades	-\$33.60	-\$38.71	-\$31.94

[1] $H_0: \ F(x) \leq G(x), \ H_1: \ F(x) > G(x)$. $F(x), \ G(x)$ = cumulative distribution functions of daily net profits from sample 1 and sample 2, respectively: p-value of 1.00 for no latency v.s. 1x latency and 1.00 for 1x latency v.s. 3x latency.

[2] HH:MM:SS.U – hours:minutes:seconds:fractions of a second.

6.3. Limit order–based strategy

The most interesting results from this paper are from the limit order strategy, where arbitrageurs mainly provide liquidity to the markets. In Table 7, we observe a gross profit of C$9.6 million with pairs of cross-listed stocks selected with supervised machine learning from the universe of 74 possible pairs (see Appendix C), and for six months of trading. Adding latency in the next columns affects the strategy's profitability by reducing the net profits by about 25%. However, the percentage of net profitable trades is rather constant between the three columns. The profitability (unprofitability) between days of trading is also quite stable. The average volume per trade is quite low and stable and is similar to that of Budish et al. [12], as can be seen in the second column of Table 5. Larger volumes with higher probability of non-execution risk could have been used, but we choose to be conservative as to minimize the impact on the price discovery process. The annual colocation cost and proprietary data feed total cost in Toronto is C$116,250 (see Table 1 for the cost breakdown). Consequently, international arbitrage of cross-listed stock is profitable with the proposed limit order strategy, even when all latencies, costs and risks are considered.

Therefore, the main question is the following: Does a net annual profit of about C$8 million (US$6 million, column 3 Table 7 with real latencies and all costs) seem reasonable for this international arbitrage activity, which can be managed by one trader in a large trading firm? Note that the original model of Budish et al. [12] with market orders generated a gross annual profit of US$8.75 million from the NYSE in 2005 (C$10.60 million), in a year where the VIX was comparable to that of 2019. But their model made only about C$2 million gross annual profits with our data in 2019, because the market activity is much less intense on the selected cross-listed stocks than on their two very liquid financial assets. Moreover, as they claim, their trading model is quite simple and they predict that a more sophisticated one should generate higher profits, which is demonstrated here in an international context with limit orders.

To remove the possibility of backtest overfitting (Bailey et al. [3]), only one set of parameters for the proposed strategies has been tested (see Appendix B). It has been applied to

Table 7. Results with our limit orders–based strategy

1	**2**	**3**	**4**
Model	**Limit Orders w/o Latency**	**Limit Orders w/ 1x Latency**	**Limit Orders w/ 3x Latency**
Latency multiplier	0	1	3
Pair selection	Yes	Yes	Yes
Gross profit	\$9,608,178.87	\$8,641,338.63	\$8,363,528.28
Loss	-\$4,757,168.60	-\$5,041,665.26	-\$5,168,902.58
Trading fees	-\$78,132.64	-\$82,067.16	-\$83,537.87
Trading rebates	\$553,201.20	\$476,071.01	\$458,542.50
Total net profit	\$5,326,078.83	\$3,993,677.22	\$3,569,630.33
Mean daily net profit	\$46,719.99	\$35,032.26	\$31,312.55
Median daily net profit	\$44,453.98	\$33,756.44	\$29,610.42
Mean daily net profit per pair, per day	\$2,273.19	\$1,704.51	\$1,523.53
$p - value$ Kolmogorov-Smirnov test[1]		1.00	1.00
1st most profitable day (date - profit)	2019/05/09 \$100,142.51	2019/05/09 \$82,330.71	2019/05/09 \$77,292.31
5th most profitable day (date - profit)	2019/05/13 \$78,509.62	2019/06/20 \$58,157.95	2019/05/07 \$53,633.28
1st most unprofitable day (date - profit)	2019/06/04 \$15,061.17	2019/03/13 \$12,210.91	2019/03/13 \$9,130.81
5th most unprofitable day (date - profit)	2019/03/18 \$22,810.62	2019/03/18 \$15,997.46	2019/03/18 \$13,349.39
Average time in trade (excl. futures contracts)	00:01:29.51	00:01:39:10	00:01:41.22
Average time in trade[2] (incl. futures contracts)	00:01:46.55	00:01:56.61	00:01:58.19
# Net profitable trades	1,063,897	930,388	892,772
# Net unprofitable trades	325,351	322,230	327,096
# Trades	1,389,248	1,252,618	1,219,868
% Net profitable trades	76.58%	74.28%	73.19%
Average volume per trade	188.10	187.99	188.36
Average net profit per trade	\$3.83	\$3.19	\$2.93
Average profit per net profitable trades	\$9.51	\$9.76	\$9.84
Average profit per net unprofitable trades	-\$14.71	-\$15.78	-\$15.94
% Trade using marketable orders	16.42%	19.56%	20.50%

[1] $H_0: F(x) \leq G(x),\ H_1: F(x) > G(x)$. $F(x),\ G(x)$ = cumulative distribution functions of daily net profits from sample 1 and sample 2, respectively: p-value of 1.00 for no latency v.s. 1x latency and 1.00 for 1x latency v.s. 3x latency.

[2] HH:MM:SS.U – hours:minutes:seconds:fractions of a second.

every pair and every day of our data. Of course, the probability that this set of parameters is optimal for any pair and any day is close to zero, and if we had backtested the strategies multiple times, we could have selected the set that generated the greatest profitability and performance metrics of our portfolio. However, by using a single set of parameters fixed before any testing, we ensure that our findings are generalizable by avoiding any overfitting behavior. Hence, the metrics that were shown in this section could be improved and the results thus offer a conservative, but reasonable measure of the profitability of international arbitrage of cross-listed stocks between Canada and the U.S..

7. Detailed limit order strategy performance

7.1. Statistics

In this section, a more detailed view of the performance of the limit order strategy in the real latency setting is presented (column 3 of Table 7). We define a captured arbitrage opportunity as an opportunity where the positions in a pair at TSX and NYSE are both opened and closed with limit orders following the arbitrage strategy described in Section 3. This excludes arbitrage opportunities where a least one leg had to be closed by the stop-loss or the time circuit breakers implemented for risk management (see Appendix B).

Figure 1 shows the mean daily number of captured arbitrage opportunities per ticker, and Figure 2 the mean duration of these trades. The number of captured arbitrage opportunities exhibits some daily fluctuations, but the quantity remains stationary over the period. On average, there are 180 captured arbitrage opportunities per ticker per day. The mean duration, computed as the mean of the daily means of captured opportunity pairs, is about 122 seconds, and is also stationary during our period of analysis. Note that both quantities are anticorrelated (Pearson correlation coefficient of -0.923). This is because the strategy does not enter a new position as long as the previous one is still open. This condition avoids building huge inventories which would involve, among other aspects, significant price impact when ending arbitrage activities. Thus, a longer time to close both legs of the strategy directly leads to a lesser number of potential arbitrage opportunities to be captured.

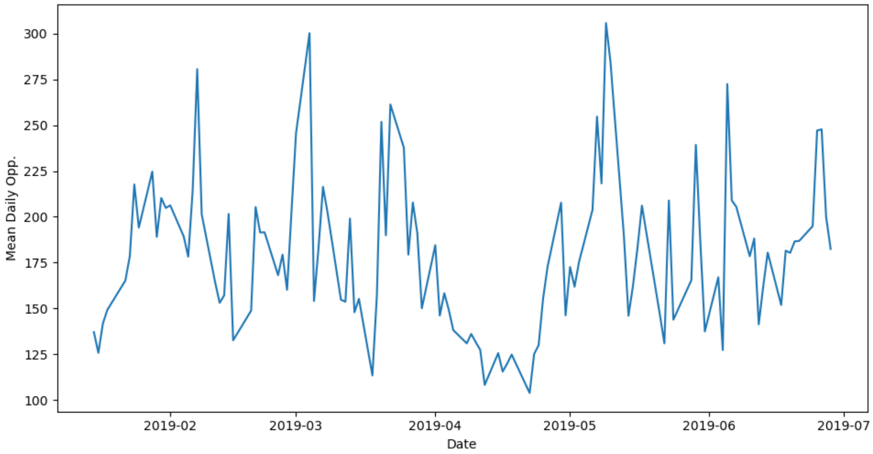

Fig. 1. Mean daily number of captured arbitrage opportunities on all selected ticker

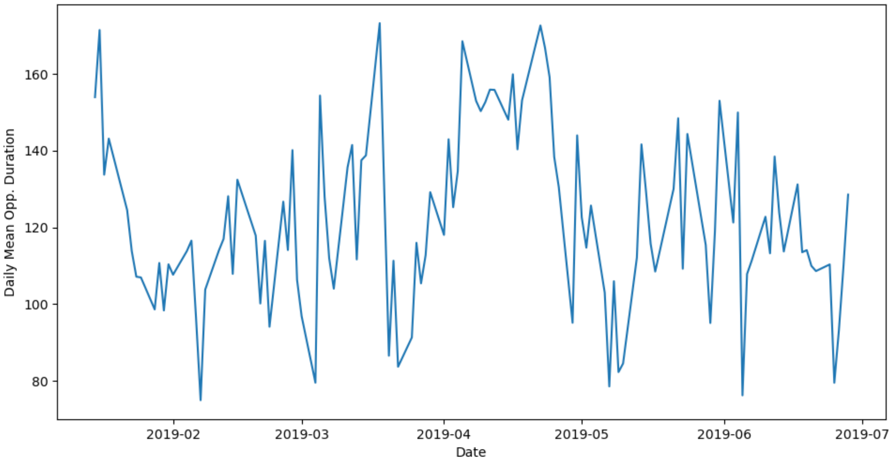

Fig. 2. Daily mean duration in seconds of captured arbitrage opportunities on all selected ticker

Figure 3 shows the daily net profit measured as the total realized net profit per day over the selected assets in the first six months of 2019, and Figure 4 as well with the average per captured arbitrage opportunity. The mean daily net profit is C$67,369 and the mean net profit per captured arbitrage opportunity is around C$19, in line with usual reported high-frequency trading activities. Per ticker, the daily mean net profit is of C$3,411.

Figure 5 shows the empirical cumulative distribution function (CDF) of the net profit per captured arbitrage opportunity in C$. Based on this CDF, 99.7% of the captured arbitrage opportunities are profitable. The median is around C$11, and the 99^{th} percentile

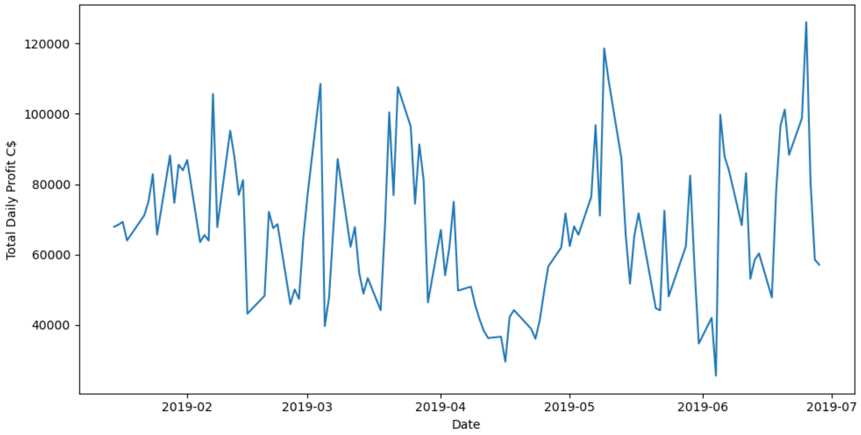

Fig. 3. Total daily net profit from captured arbitrage opportunities on all selected tickers

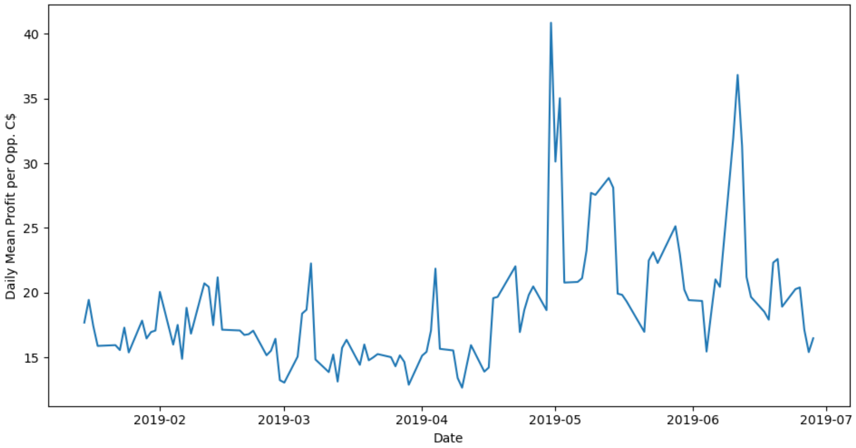

Fig. 4. Daily mean profitability of captured arbitrage opportunities on all selected tickers

is around C$110. This confirms the theoretical validity of the strategy, meaning that when an arbitrage opportunity is perfectly captured with limit orders, it is almost guaranteed to be profitable. The remaining 0.3% of unprofitable captured arbitrage opportunities exist because the positions cannot always close at the exact equilibrium value, as explained in Appendix B.

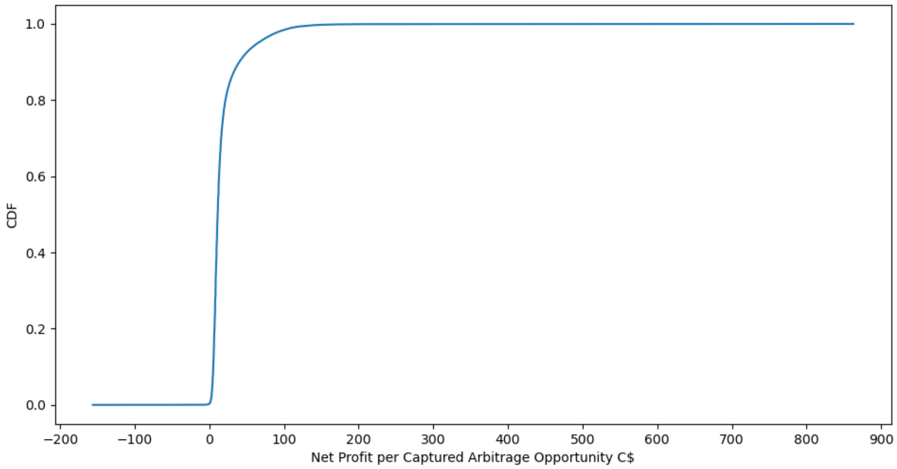

Fig. 5. Empirical CDF of net profit per captured arbitrage opportunity

7.2. Regression analysis

We employ a regression analysis to better understand the stylized facts affecting the daily net profitability of the strategy. Using standard variables such as the intraday mid-price volatility of the assets traded at exchange $k \in \{TSX, NYSE, CME\}$ on day d ($vol_{k,d}$), the average bid-ask spreads ($spread_{k,d}$), the total trading volumes ($trade_{k,d}$), and the total quantity of messages resulting from L1 updates ($messages_{k,d}$), all in their respective currency, we want to explain the average net profitability of the selected pairs on day d ($\overline{profit}_d$). Every variable is computed as the weighted mean of the stock–level variable in the selected pairs on day d, where the weight given to a specific stock is the proportion of its daily traded value, compared with the total traded value for every stock of the same exchange in our portfolio on that day (all in C\$). Table 8 reports the descriptive statistics of these variables. In Appendix A, all variables are described in Table 14, and Table 15 reports their Pearson correlation coefficients.

The mid-price volatilities of cross-listed stocks have similar distributions on both stock exchanges. The same applies for the spread and the number of messages from LOB level one. On the other hand, the volume of trades at TSX is almost three times greater than at NYSE, which is expected from a portfolio composed entirely of Canadian stocks.

Table 8. Descriptive statistics of variables used in the regression analysis to explain the daily net profit of the strategy with limit orders

Variable	Mean	Std. Dev.	Min.	Q1	Median	Q3	Max.
$\overline{profits}$	3,411	1237	1,636	2,543	3,201	4,002	8,471
vol_{TSX}	0.458	0.143	0.259	0.361	0.412	0.524	0.974
vol_{NYSE}	0.467	0.151	0.269	0.357	0.420	0.548	1.007
vol_{CME}	0.086	0.047	0.024	0.054	0.074	0.114	0.244
$spread_{TSX}$	5.791	1.267	3.567	4.916	5.688	6.213	1.097
$spread_{NYSE}$	6.854	1.279	4.800	5.835	6.732	7.385	1.079
$spread_{CME}$	0.576	0.020	0.542	0.566	0.576	0.584	0.715
$trade_{TSX}$	775,033	361,920	349,538	506,020	686,478	949,768	2,288,334
$trade_{NYSE}$	280,935	153,312	118,714	183,374	226,571	296,655	973,461
$trade_{CME}$	64,664	75,053	4,195	27,383	34,537	46,817	297,363
$messages_{TSX}$	59,136	13,474	35,229	48,619	58,494	67,034	94,736
$messages_{NYSE}$	53,719	13,253	32,036	43,001	51,791	62,492	101,549
$messages_{CME}$	192,958	55,922	66,246	150,944	186,874	223,187	346,698
Count	114						

From Table 15 of Appendix A, a significant and positive relationship between the strategy's profitability and the volatility of the markets can be observed. The stocks' bid-ask spread is the most highly and positively correlated variable with the profitability of the strategy, which is expected since the strategy uses limit orders. Finally, all the messages variables are statistically and positively correlated with the strategy's profitability, which will be explained later in this section.

As expected, pairs of the same variables on TSX and NYSE are highly correlated. To reduce potential multicollinearity, each of these is combined into one variable by using the mean of the respective TSX and NYSE values, thus creating the variables $\overline{vol}_{stocks,d}$, $\overline{spread}_{stocks,d}$, $\overline{trade}_{stocks,d}$ and $\overline{messages}_{stocks,d}$. The linear regression model is written as follows, for day $d \in \{1,2,\ldots,114\}$:

$$\begin{aligned}\overline{profits}_d =& \beta_0 + \beta_1 vol_{CME,d} + \beta_2 \overline{vol}_{stocks,d} + \beta_3 spread_{CME,d} + \beta_4 \overline{spread}_{stocks,d} \\ &+ \beta_5 trade_{CME,d} + \beta_6 \overline{trade}_{stocks,d} + \beta_7 messages_{CME,d} + \beta_8 \overline{messages}_{stocks,d} + \varepsilon_d\end{aligned}$$

where $\varepsilon_d \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),\ \forall d$. The regression coefficients are obtained by ordinary least squares (OLS), and the covariance matrix is estimated with the heteroskedasticity and autocorrelation consistent approach of Newey and West [46]. Table 9 summarizes the regression results. As it suggests, the number of L1 update messages, the size of the spread and the trading volume of the stocks all contribute significantly to the daily net profits generated by the portfolio of cross-listed stock pairs. These results are consistent with the machine learning pair selection methodology (see Appendix C for more details). A larger spread for the stocks is directly beneficial to the limit order strategy, which can be explained by equations (4), (5) and (6). Together, these equations demonstrate that a larger spread leads to a higher profit for any given arbitrage opportunity, and that the profitable arbitrage opportunities are more frequent for days with larger spreads. As for the number of messages, the result is intuitive because a higher L1 activity generally increases the likelihood of active limit orders

Table 9. OLS linear regression for the average daily net profitability of the limit order strategy with Newey-West covariance matrix estimation.

Variable	Coefficient	p-value
β_0	-3,823.011	0.202
vol_{CME}	-2,533.717	0.103
$\overline{vol}_{stocks}$	695.229	0.284
$spread_{CME}$	-1,817.241	0.723
$\overline{spread}_{stocks}$	791.142	0.000
$trade_{CME}$	0.001	0.518
$\overline{trade}_{stocks}$	-0.002	0.009
$messages_{CME}$	0.003	0.139
$\overline{messages}_{stocks}$	0.069	0.000
$Adj.\ R^2$	0.662	
$F-stat$	22.570	

to be filled, or canceled by the risk management circuit breakers in the case where assets' prices deviate from limit orders' prices. Hence, the more messages are observed, the faster the orders can be executed or canceled, and the faster the strategy can move on to the next opportunity (which was observed in Figures 1 and 2), as opposed to days when markets are quieter and limit orders can remain in the LOB for longer periods of time. Lastly, a larger trading volume contributes negatively to profitability, especially at the NYSE. The higher latency to that exchange prevents the strategy from reacting very rapidly compared to other participants colocated there. Thus, trades occurring before its limit orders are included in the LOB (or even before the information was analyzed by the algorithm) can cause the mispricing to dissipate.

7.3. Macroeconomic environment effects

The goal of this section is to provide a robustness analysis of our model in different macroeconomic environments by comparing its profitability across our data. Table 10 analyzes the strategy's results for the six months of 2019 where statistically different stock-return distributions occurred. We use first-order stochastic dominance to rank the monthly stock-return' distributions. Stochastic dominance quantifies whether one probability distribution is greater than another. Given two distributions F and G, it is said that F has a first-order stochastic dominance over G if and only if $F(x) \leq G(x)$, $\forall x \in \mathbb{R}$ with strict inequality for some x. One popular test for first-order stochastic dominance is the one-sided Kolmogorov-Smirnov test (Schmid and Trede [53]). We apply this test to every pair of monthly returns and order them from most dominant (rank 1) to least dominant (rank 6). Hence, rank 1 is the month with the statistically highest return distribution. Some pairs of monthly return' distributions cannot reject the two-sided Kolmogorov-Smirnov null hypothesis that the two distributions are identical, so these pairs are of equal rank.

From Table 10, the performance of the strategy can be analyzed in two market extremes, namely, in the great uptrend market of January 2019 (rank 1) and in the considerable downtrend market of May 2019 (rank 6). In both months, the strategy fares well, but markedly so in May, where it generated the greatest average net daily profit out of the entire data sample. In down markets like May, bid-ask spreads increase (Chordia et al. [17]), which is an advantage for our strategy, as shown in the previous section. This fact can be observed in the average net profit per stock traded, which, in May, is almost triple that of January. But liquidity severely decreases during downtrend markets, as opposed to uptrend markets (Chordia et al. [17]), so the volume per arbitrage opportunity is significantly greater in January than in May, which counterbalances narrower bid-ask spreads and still results in a net profit. In more regular macroeconomic environments, e.g., February, March, April, and June 2019, the strategy remains profitable. The strategy is market neutral, so it should remain applicable in any macroeconomic environment, as Table 10 suggests.

Table 10. Monthly statistics of cross-listed stocks returns, and the strategy's respective statistics

Statistic	**January**	**February**	**March**	**April**	**May**	**June**
Avg. Returns[1]	10.25%	3.01%	0.89%	1.17%	-6.60%	3.67%
Std. Returns	12.51%	6.97%	6.97%	8.05%	10.45%	9.86%
Stochastic Dominance Rank[2]	1	2	4	4	6	2
Avg. Daily Net Profit	\$31,104.08	\$29,753.21	\$28,712.23	\$22,189.34	\$40,215.74	\$37,382.49
Avg. Daily # of Trades	12,834	11,652	13,711	8,381	9,004	11,206
Avg. Volume Per Trade	199.59	186.45	201.19	164.28	144.28	160.23
Avg. Net Profit Per Stock Traded	\$0.0121	\$0.0137	\$0.0104	\$0.0161	\$0.0310	\$0.0208

[1] Monthly returns are computed as the return from first to last trade price occurring in the specified month.

[2] The stochastic dominance ranking, from most dominant to less dominant. Found by applying the one-sided, two-sample Kolmogorov-Smirnov test at a p-value of 1% to every pair of monthly return distributions. The two-sided, two-sample Kolmogorov-Smirnov test was also applied to confirm when pairs had no statistically verified stochastic dominance from the one-sided Kolmogorov-Smirnov test.

7.4. Profitability

Figure 6 shows the net cumulated profits over the entire period on a trade basis. There is minimal intraday drawdown, and as was shown in Figure 3, the net daily profits are stationary, which explains the quasi linearity of the function in Figure 6.

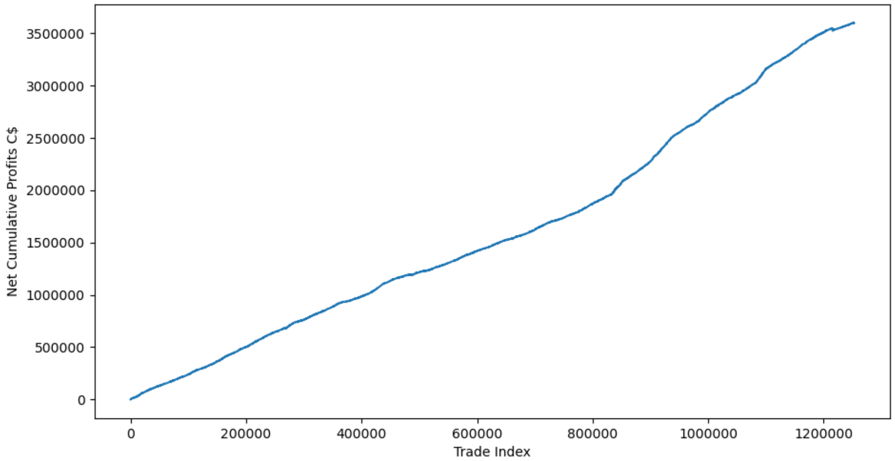

Fig. 6. Net cumulated profits in C$ on a trade basis over the entire period

Figure 7 presents the daily maximum net aggregated positions taken at each exchange for our portfolio of selected pairs. The maximum net open position in absolute value is around C$453,000 at TSX, C$465,000 at NYSE, and C$590,000 at CME, meaning that an investment of C$1,000,000 to cover the margins is more than enough. Note that only a margin of US$1,100 per CAD/USD futures contract is needed at the CME. Given the annual net profit of C$8 million generated by the strategy in 2019, this results in an annual net return of 700%. Figure 8 shows the empirical CDF of the needed aggregated net margin in C$. This margin at time t, M_t, can be expressed as follows:

$$M_t = |V_{TSX,t}| + \left|V_{NYSE,t} - \frac{G_{NYSE,t}}{r_t}\right| + \frac{1{,}100}{r_t}\left|V_{CME,t} - \frac{G_{CME,t}}{r_t}\right|/100{,}000,$$

where $V_{TSX,t}$, $V_{NYSE,t}$ and $V_{CME,t}$ are the portfolio exposures in C$ in the respective exchanges introduced in subsection 3.5. Once again, it can be seen that a capital of C$1,000,000 always covers the margins in the three exchanges, while C$185,000 covers 80% of the needed margins at any time, meaning that the high levels of aggregated positions are transitory.

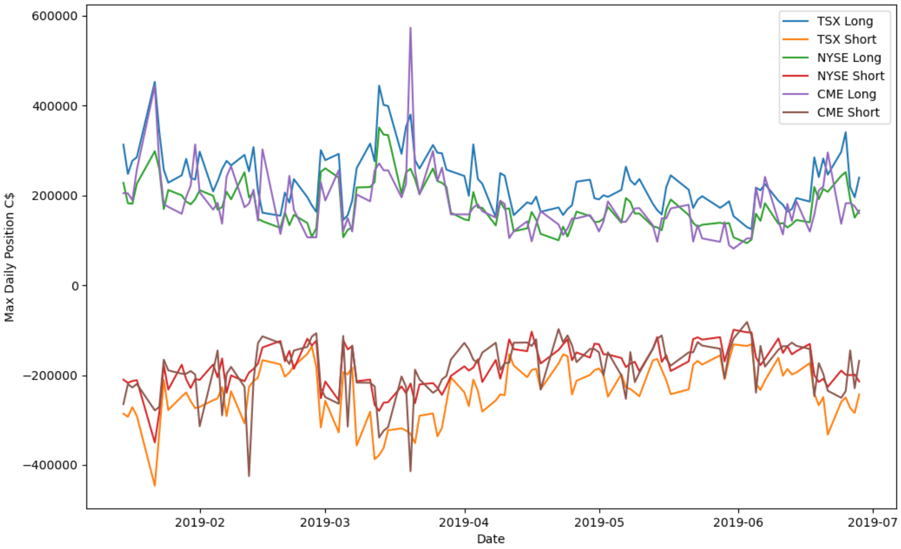

Fig. 7. Maximum daily net aggregated long and short positions of the selected pairs portfolio at the three exchanges

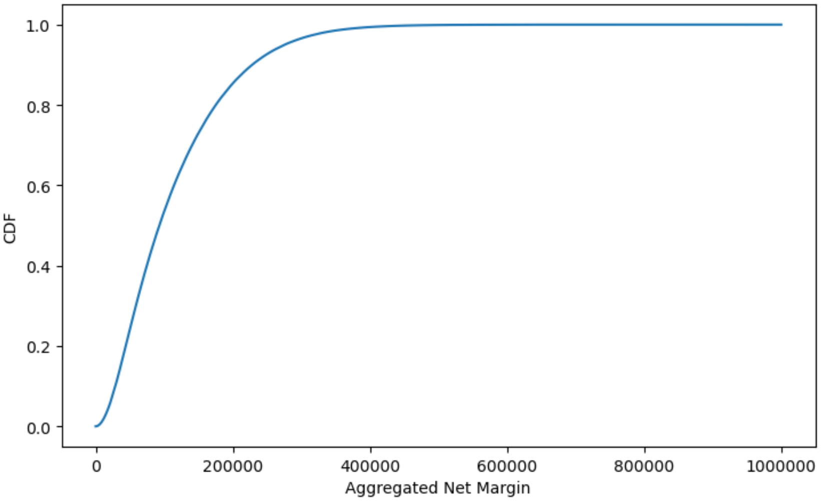

Fig. 8. Empirical CDF of the needed aggregated net margin in C$

The annualized Sharpe ratio computed from the daily returns and the margin of C$1 million is 51.04. It is very high, but the daily profits are perfectly comparable to the trading profits of HFTers found in Baron et al. [5]. This result is explained by the low volatility of the profits as seen in Figures 3 and 5. Also the Deflated Sharpe Ratio proposed by Bailey and de Prado [4] is approximately equal to 1. This high value results from the fact that there is only one backtesting trial, hence there is no variance across the trials and a quasi-null likelihood of a false discovery.

8. Conclusion

The profitability of high-frequency arbitrage activities in international cross-listed stocks is studied with a novel trading strategy generalizable to a broad cross-listed assets universe. The theoretical strategy signals when the prices of cross-listed stocks deviate enough from their relative equilibrium that an economically viable arbitrage opportunity occurs. The model is applied to North American markets during the first six months of 2019, namely to the New York Stock Exchange (NYSE) and the Chicago Mercantile Exchange (CME) in the United States, and the Toronto Stock Exchange (TSX) in Canada.

This paper is the first to examine stocks' cross-country mean-reverting arbitrage. The work is based on a unique temporal frame of reference, meaning the data feeds from all exchange venues are synchronized by explicitly taking into account the latency that comes from the transmission of information between them and the information processing time. All potential arbitrage trading costs are also considered. The profits obtained by the limit order strategy are reasonable when compared with previous contributions in the literature. But international arbitrage with market orders is not profitable on our data, because of the great interconnectedness between Canadian and American exchanges. We also show how the profitability of high-frequency arbitrage is often overestimated in previous studies by not considering both the practical aspects of trading and market frictions.

The original goal was not to contribute to the normative discussion about the effect of continuous HFT on the general welfare of financial markets. Rather, it was to replicate the precise behavior of a high-frequency trading firm as to provide a better understanding of their arbitrage activities. This research highlights the high-frequency arbitrageur's economic incentive to act as a liquidity provider, and the importance of considering real market frictions in HFT research. These results could be useful to improve the understanding of high-frequency trading's complex and secretive nature. The proposed model can be deployed in a real-time environment by institutional investors, professional arbitrageurs, market makers, hedgers, and regulators. Our approach provides a contemporary understanding of an economically viable arbitrage approach that helps restore equilibrium in financial markets. These arbitrage activities are usually carried out by the largest traders under strong competition. They provide liquidity to the markets and are remunerated for this activity. Are the profits they earn too high? The results of this study do not provide a conclusive answer to this question, but we have demonstrated that high-frequency traders can make sizable arbitrage profits under fair trading conditions.

Finally, do these arbitrage activities affect long-term investors, who are not involved in arbitrage activities, which represent most stock investors? We do not have sufficient data to answer this question, and this issue warrants additional quantitative research.

Acknowledgments

Financial support from the Canada Research Chair in Risk Management, Canada First Research Excellence Fund, and IVADO is graciously acknowledged.

References

[1] Ait-Sahalia, Y. and Saglam, M. (2023). High frequency market making: The role of speed. *Journal of Econometrics*. In press, DOI: https://doi.org/10.1016/j.jeconom.2022.12.015.

[2] Aquilina, M., Budish, E., and O'Neill, P. (2022). Quantifying the high-frequency trading 'arms race'. *The Quarterly Journal of Economics*, 137:493–564.

[3] Bailey, D., Borwein, J., de Prado, M. L., and Zhu, Q. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61:458–471.

[4] Bailey, D. and de Prado, M. L. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40:94–107.

[5] Baron, M., Brogaard, J., and Kirilenko, A. (2012). The trading profits of high frequency traders. Working paper, Princeton University, University of Washington, Massachusetts Institute of Technology.

[6] Biais, B., Foucault, T., and Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics*, 116:292–313.

[7] Biais, B. and Woolley, P. (2011). High frequency trading. Working paper, Toulouse School of Economics, University of Toulouse.

[8] Blume, M. and Goldstein, M. (1991). Differences in execution prices among the NYSE, the regionals, and the NASD. *Rodney L. White Center for Financial Research Paper*, pages 4–92.

[9] Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., and Sokolov, K. (2018). High frequency trading and extreme price movements. *Journal of Financial Economics*, 128:253–265.

[10] Brolley, M. (2020). Price improvement and execution risk in lit and dard markets. *Management Science*, 66:863–886.

[11] Brunnermeier, M. and Pedersen, L. (2008). Market liquidity and funding liquidity. *The Review of Financial Studies*, 22:2201–2238.

[12] Budish, E., Cramton, P., and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quaterly Journal of Economics*, 130(4):1547–1621.

[13] Campbell, R. and Huang, R. (1991). Volatility in the foreign currency futures market. *The Review of Financial Studies*, 32:1068–1101.

[14] Chao, Y., Chen, Y., and Mao, Y. (2019). Why discrete price fragments U.S. stock exchanges and disperses their fee structures. *The Review of Financial Studies*, 32:1068–1101.

[15] Chen, H., Chen, S., Chen, Z., and Li, F. (2019a). Empirical investigation of an equity pairs trading strategy. *Management Science*, 65:370–389.

[13] Chen, Y., Da, Z., and Huang, D. (2019b). Arbitrage trading: The long and the short of it. *The Review of Financial Studies*, 32:1608–1646.

[17] Chordia, T., Roll, R., and Subrahmanyam, A. (2001). Market liquidity and trading activity. *The Journal of Finance*, 55:501–530.

[18] Chouinard, E. and D'Souza, C. (2003). The rationale for cross-border listings. *Bank of Canada Review*, Winter:23–30.

[19] Dahlström, H. and Nordén, L. (2018). The determinants of limit order cancellations. Mimeo, Stockholm Business School.

[20] Dewhurst, D., Oort, C. V., IV, J. R., Gray, T., Danforth, C., and Tivnan, B. (2019). Scaling of inefficiencies in the U.S. equity markets: Evidence from three market indices and more than 2900 securities. Available at https://arxiv.org/abs/1902.04691.

[21] Ding, S., Hanna, J., and Hendershott, T. (2014). How slow is the NBBO? A comparison with direct exchange feeds. *The Financial Review*, 49:313–332.

[22] Dixon, M., Polson, N., and Sokolov, V. (2019). Deep learning for spatio-temporal modeling: Dynamic traffic flows and high frequency trading. *Applied Stochastic Models in Business and Industry*, 35:788–807.

[23] Dugast, J. (2018). Unscheduled news and market dynamics. *The Journal of Finance*, 73:2537–2586.

[24] Engelberg, J., Pengjie, G., and Jagannathan, R. (2009). An anatomy of pairs trading: The role of idiosyncratic news, common information and liquidity. *Third Singapore International Conference on Finance.*

[25] Eun, C. and Sabherwal, S. (2003). Cross-border listings and price discovery: Evidence from U.S.-listed Canadian stocks. *The Journal of Finance*, 58:549–575.

[26] Foucault, T. and Biais, B. (2014). HFT and market quality. *Bankers, Markets & Investors*, 128:5–19.

[27] Foucault, T., Kozhan, R., and Tham, W. (2017). Toxic arbitrage. *The Review of Financial Studies*, 30:1053–1094.

[28] Foucault, T. and Moinas, S. (2019). Is trading fast dangerous? In Mattli, W., editor, *Global Algorithmic Capital Markets: High Frequency Trading, Dark Pools, and Regulatory Challenges*, chapter 2, pages 9–27. Oxford University Press.

[29] Frazzini, A., Israel, R., and Moskowitz, T. (2018). Trading costs. Available at https://papers.ssrn.com/abstract_id=3229719.

[30] Friederich, S. and Payne, R. (2015). Order-to-trade ratios and market liquidity. *Journal of Banking & Finance*, 50:214–223.

[31] Gagnon, L. and Karolyi, G. (2013). Chapter 11 - International Cross-listings. In Caprio, G., Beck, T., Claessens, S., and Schmukler, S. L., editors, *The Evidence and Impact of Financial Globalization*, pages 155–180. Academic Press, San Diego.

[32] Garriott, C., Pomeranets, A., Slive, J., and Thorn, T. (2013). Fragmentation in Canadian equity markets. White Paper, Bank of Canada Review.

[33] Gatev, E., Goetzmann, W., and Rouwenhorst, K. (2006). Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19:797–827.

[34] Goldstein, M., Kumar, F., and Graves, F. (2014). Computerized and high frequency trading. *The Financial Review*, 49:177–202.

[35] Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, 50:1175–1199.

[36] Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16:646–679.

[37] Hendershott, T., Jones, C., and Menkveld, A. (2011). Does algorithmic trading improve liquidity? *The Journal of Finance*, 66:1–33.

[38] Jones, C. (2013). What do we know about high-frequency trading? Columbia Business School Research Paper. No. 13-11.

[39] Kozhan, R. and Tham, W. (2012). Execution risk in high-frequency arbitrage. WBS Finance Group Research Paper. No. 179.

[40] Krauss, C. (2017). Statistical arbitrage pairs trading strategies: Review and outlook. *Journal of Economic Surveys*, 31:513–545.

[31] Lee, C. (1993). Market integration and price execution for NYSE–listed securities. *The Journal of Finance*, 48:1009–1038.

[42] Liu, W. (2009). Monitoring and limit order submission risks. *Journal of Financial Markets*, 12:107–141.

[43] Mavroudis, V. (2019). Market manipulation as a security problem. Working paper, University College London.

[44] Menkveld, A. (2014). High-frequency traders and market structure. *The Financial Review*, 49:333–344.

[45] Menkveld, A. (2016). The economics of high-frequency trading: Taking stock. *Annual Review of Financial Economics*, 8:1–24.

[46] Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

[47] O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economic*, 116:257–270.

[48] Pontiff, J. (2006). Costly arbitrage and the myth of idiosyncratic risk. *Journal of Accounting and Economics*, 42:35–52.

[49] Rein, C., Rüschendorf, L., and Schmidt, T. (2021). Generalized statistical arbitrage concepts and related gain strategies. *Mathematical Finance*, 31:563–594.

[50] Reuters, T. (2020). Competition to heat up among U.S. stock exchanges with new entrants. Available at https://www.reuters.com/article/us-usa-exchanges-idUSKBN25H23K. Accessed February 25, 2021.

[51] Riordan, R. and Storkenmaier, A. (2012). Liquidity and price discovery. *Journal of Financial Markets*, 15:416–437.

[52] Scherrer, C. (2018). Information processing on equity prices and exchange rate for cross-listed stocks. *Journal of Financial Markets*, 54:100634.

[53] Schmid, F. and Trede, M. (1996). Testing for first-order stochastic dominance: A new distribution-free test. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45:371–380.

[54] Shkilko, A., Ness, B. V., and Ness, R. V. (2008). Locked and crossed markets on NASDAQ and the NYSE. *Journal of Financial Markets*, 11:308–337.

[55] Shkilko, A. and Sokolov, K. (2020). Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. *The Journal of Finance*, 75:2899–2927.

[56] Spraos, J. (1953). The theory of forward exchange and recent practice. *The Manchester Schools*, 21:87–117.

[56] Stübinger, J. and Bredthauer, J. (2017). Statistical arbitrage pairs trading with high-frequency data. *International Journal of Economics and Financial Issues*, 7:650–662.

[58] Tivnan, B., Dewhurst, D., Oort, C. V., IV, J. R., Gray, T., Koehler, M., McMahon, M., Slater, D., Veneman, J., and Danforth, C. (2020). Fragmentation and inefficiencies in US equity markets: Evidence from the Dow 30. *PLOS ONE*, 15(1):1–24.

[59] TSX (2018). TSX trading fee schedule. Available at https://www.tsx.com/resource/en/1756/tsx-trading-fee-schedule-effective-june-4-2018-en.pdf. Accessed February 25, 2021.

[60] Wah, E. (2016). How prevalent and profitable are latency arbitrage opportunities on U.S. stock exchanges? Available at https://ssrn.com/abstract=2729109.

Appendix A. Additional tables

Table 11. List of available cross-listed stocks with the TSX ticker and NYSE ticker counterpart. Also included are the company's name, economic sector and S&P/TSX 60 membership status.

TSX Ticker	NYSE Ticker	Company	Sector	S&P/TSX 60
ABX	GOLD	Barrick Gold Corp.	Materials	Yes
AEM	AEM	Agnico Eagle Mines Ltd.	Materials	Yes
AGI	AGI	Alamos Gold Inc.	Mining	No
AQN	AQN	Algonquin Power & Utilities Corp.	Clean Technology	No
ATP	AT	Atlantic Power Corp.	Clean Technology	No
BAM.A	BAM	Brookfield Asset Management Inc.	Financials	Yes
BB	BB	Blackberry Ltd.	Information Technology	Yes
BCB	COT	Cott Corp.	Consumer Products & Services	No
BCE	BCE	BCE Inc.	Telecommunication Services	Yes
BMO	BMO	Bank of Montreal	Financials	Yes
BNS	BNS	Bank of Nova Scotia	Financials	Yes
BTE	BTE	Baytex Energy Corp.	Oil & Gas	No
BXE	BXE	Bellatrix Exploration Ltd.	Oil & Gas	No
CAE	CAE	CAE Inc.	Technology	No
CCO	CCJ	Cameco Corp.	Energy	Yes
CLS	CLS	Celestia Inc.	Technology	No
CM	CM	Canadian Imperial Bank of Commerce	Financials	Yes
CNQ	CNQ	Canadian Natural Resources Ltd.	Energy	Yes
CNR	CNI	Canadian National Railway Company	Industrials	Yes
CNU	CEO	CNOOC Ltd.	Oil & Gas	No
CP	CP	Canadian Pacific Railway Ltd.	Industrials	Yes
CPG	CPG	Crescent Point Energy Corp.	Energy	Yes
CVE	CVE	Cenovus Energy Inc.	Energy	Yes
ECA	ECA	Encana Corp.	Energy	Yes
EDR	EXK	Endeavour Silver Corp.	Mining	No
ELD	EGO	Eldorado Gold Corp.	Mining	No
ENB	ENB	Enbridge Inc.	Energy	Yes
ERF	ERF	Enerplus Corp.	Oil & Gas	No
FNV	FNV	Franco-Nevada Corp.	Materials	Yes
FR	AG	First Majestic Silver Corp.	Mining	No
FTS	FTS	Fortis Inc.	Utilities	Yes
FVI	FSM	Fortuna Silver Mines Inc.	Mining	No

Table 11 continued from previous page

TSX Ticker	NYSE Ticker	Company	Sector	S&P/TSX 60
G	GG	Goldcorp Inc.	Materials	Yes
GIB.A	GIB	CGI Group Inc.	Information Technology	Yes
GIL	GIL	Gildan Activewear Inc.	Consumer Discretionary	Yes
GOOS	GOOS	Canada Goose Holdings Inc.	Consumer Products & Services	No
HBM	HBM	HudBay Minerals Inc.	Mining	No
IMG	IAG	IAMGold Corp.	Mining	No
JE	JE	Just Energy Group Inc.	Utilities & Pipelines	No
K	KGC	Kinross Gold Corp.	Materials	Yes
KL	KL	Kirkland Lake Gold Ltd.	Mining	No
LAC	LAC	Lithium Americas Corp.	Mining	No
MFC	MFC	Manulife Financial Corp.	Financials	Yes
MG	MGA	Magna International Inc.	Consumer Discretionary	Yes
NEXA	NEXA	Nexa Resources S.A.	Mining	No
NOA	NOA	North American Construction Group	Industrial Products & Services	No
OR	OR	Osisko Gold Royalties Ltd.	Mining	No
OSB	OSB	Nordbord Inc.	Industrial Products & Services	No
PD	PDS	Precision Drilling Corp.	Industrial Products & Services	No
PPL	PBA	Pembina Pipeline Corp.	Utilities	Yes
PVG	PVG	Pretium Resources Inc.	Mining	No
QSR	QSR	Restaurant Brands International Inc.	Consumer Discretionary	Yes
RBA	RBA	Ritchies Bros. Auctioneers Inc.	Industrial Products & Services	No
RCI.B	RCI	Rogers Communication Inc.	Telecommunication Services	Yes
RFP	RFP	Resolute Forest Products Inc.	Industrial Products & Services	No
SEA	SA	Seabridge Gold Inc.	Mining	No
SHOP	SHOP	Shopify Inc.	Technology	No
SJR.B	SJR	Shaw Communications Inc.	Telecommunication Services	Yes
SLF	SLF	Sun Life Financials Inc.	Financials	Yes
STN	STN	Stantec Inc.	Industrial Products & Services	No
SU	SU	Suncor Energy Inc.	Energy	Yes
T	TU	Telus Corp.	Telecommunication Services	Yes
TA	TAC	TransAlta Corp.	Utilities & Pipelines	Yes
TD	TD	Toronto-Dominion Bank	Financials	Yes
TECK.B	TECK	Teck Resources Ltd.	Materials	Yes
THO	TAHO	Tahoe Resources Inc.	Mining	No
TRI	TRI	Thomson Reuters Corp.	Consumer Discretionary	Yes
TRP	TRP	TransCanada Corp.	Energy	Yes
TRQ	TRQ	Turquoise Hill Resources Ltd.	Mining	No

Table 11 continued from previous page

TSX Ticker	NYSE Ticker	Company	Sector	S&P/TSX 60
UFS	UFS	Domtar Corp.	Consumer Products & Services	No
VET	VET	Vermilion Energy Inc.	Oil & Gas	No
WEED	CGC	Canopy Growth Corp.	Life Sciences	No
WPM	WPM	Wheaton Precious Metals Corp.	Mining	No
YRI	AUY	Yamana Gold Inc.	Mining	No

Table 12. Aggregated statistics of every pair from January 7, 2019 to June 28, 2019. # Trades is the total number of trades. # Quotes is the total number of quote messages. Volatility is the annualized standard deviation of the daily returns computed from close prices. Min., Mean, Median and Max. prices are respectively the minimum, empirical mean, empirical median and maximum trade prices during that period. All prices are in local currency.

TSX Ticker | NYSE Ticker	**# Trades**	**# Quotes**	**Volatility**	**Min. Price**	**Mean Price**	**Median Price**	**Max. Price**
ABX | GOLD	1355455 | 411552	12609016 | 16754385	0.3 | 0.31	15.37 | 11.52	20.63 | 15.77	20.64 | 15.77	21.67 | 16.44
AEM | AEM	712143 | 245409	11968871 | 9549364	0.23 | 0.24	51.39 | 38.72	66.73 | 50.99	66.73 | 50.99	69.13 | 52.5
AGI | AGI	342922 | 114855	3003641 | 7242356	0.4 | 0.42	4.89 | 3.68	7.82 | 5.98	7.83 | 5.99	8.25 | 6.27
AQN | AQN	454837 | 46109	3530170 | 2506458	0.12 | 0.12	13.5 | 10.13	15.87 | 12.13	15.88 | 12.13	16.6 | 12.54
ATP | AT	33558 | 23008	373750 | 653362	0.34 | 0.35	2.97 | 2.23	3.13 | 2.39	3.13 | 2.39	4.01 | 3.01
BAM.A | BAM	690629 | 252633	12662244 | 15994647	0.12 | 0.13	52.49 | 39.35	62.41 | 47.68	62.4 | 47.68	65.06 | 48.72
BB | BB	482209 | 154435	4813106 | 6724215	0.36 | 0.37	9.31 | 7.1	9.74 | 7.44	9.75 | 7.45	13.74 | 10.29
BCB | COT	171235 | 115407	3257388 | 5107512	0.28 | 0.26	16.9 | 12.73	17.35 | 13.27	17.33 | 13.27	21.06 | 15.92
BCE | BCE	829443 | 142454	7381944 | 12112698	0.1 | 0.11	53.05 | 39.75	59.47 | 45.43	59.5 | 45.44	62.75 | 47.14
BMO | BMO	927212 | 141737	8251253 | 9050862	0.11 | 0.13	88.92 | 66.65	98.76 | 75.45	98.76 | 75.45	106.51 | 79.34
BNS | BNS	1170563 | 151151	11897780 | 16543643	0.1 | 0.12	68.29 | 50.58	70.23 | 54.28	70.22 | 54.27	75.93 | 57.61
BTE | BTE	351485 | 45578	2114183 | 3694628	0.58 | 0.59	1.9 | 1.42	2.01 | 1.54	2.02 | 1.54	3.13 | 2.32
BXE | BXE	4760 | 724	18630 | 59042	0.87 | 0.74	0.48 | 0.35	0.62 | 0.47	0.62 | 0.46	0.74 | 0.55
CAE | CAE	357585 | 42689	2499568 | 2888853	0.27 | 0.27	24.99 | 18.74	35.07 | 26.79	35.1 | 26.81	36.86 | 27.42
CCO | CCJ	457592 | 139670	4065374 | 13566637	0.24 | 0.25	13.42 | 9.92	13.86 | 10.6	13.9 | 10.62	17.12 | 13.03
CLS | CLS	163847 | 61588	1530011 | 2468489	0.42 | 0.42	8.26 | 6.17	8.93 | 6.82	8.93 | 6.83	13.08 | 9.96
CM | CM	898170 | 112553	9389778 | 8047186	0.14 | 0.16	12.85 | 74.37	102.67 | 78.44	102.67 | 78.45	115.07 | 87.35
CNQ | CNQ	1679058 | 376524	17383323 | 37749573	0.27 | 0.29	33.76 | 25.33	35.13 | 26.85	35.15 | 26.86	42.56 | 31.76
CNR | CNI	885761 | 207348	7133075 | 6098839	0.13 | 0.14	100.34 | 75.18	120.61 | 92.12	120.64 | 92.11	127.96 | 95.08
CNU | CEO	156 | 41310	1936032 | 1056232	0.23 | 0.25	203.96 | 154.96	223.75 | 170.96	223.7 | 170.85	255.18 | 193.52
CP | CP	294566 | 130121	3407525 | 4345765	0.16 | 0.17	239.65 | 179.68	306.01 | 233.77	305.76 | 233.56	318.75 | 241.2
CPG | CPG	726046 | 101013	4085865 | 8280375	0.6 | 0.62	3.24 | 2.44	4.31 | 3.3	4.32 | 3.3	5.98 | 4.45

Table 12 continued from previous page

TSX Ticker | NYSE Ticker	# Trades	# Quotes	Volatility	Min. Price	Mean Price	Median Price	Max. Price
CVE | CVE	1125554 | 186950	8954509 | 17429240	0.33 | 0.35	9.62 | 7.24	11.46 | 8.75	11.46 | 8.75	14.26 | 10.6
ECA | ECA	1349297 | 395043	10526262 | 20763317	0.43 | 0.45	6.12 | 4.56	6.61 | 5.05	6.61 | 5.05	10.35 | 7.7
EDR | EXK	77642 | 47413	785259 | 2389713	0.46 | 0.47	2.27 | 1.69	2.69 | 2.06	2.7 | 2.06	3.84 | 2.85
ELD | EGO	376622 | 117803	3826242 | 6586161	0.74 | 0.76	3.36 | 2.52	7.55 | 5.77	7.55 | 5.77	7.65 | 5.82
ENB | ENB	1838436 | 310091	15384861 | 21529655	0.18 | 0.18	43.74 | 32.76	46.92 | 35.85	46.92 | 35.85	51.22 | 38.04
ERF | ERF	458784 | 111965	4590223 | 10576424	0.38 | 0.4	8.76 | 6.54	9.88 | 7.54	9.86 | 7.54	13.1 | 9.73
FNV | FNV	440300 | 136221	4644852 | 4540514	0.2 | 0.2	90.2 | 67.97	110.48 | 84.38	110.49 | 84.41	114.36 | 86.87
FR | AG	267277 | 133209	3895094 | 10620884	0.43 | 0.44	6.67 | 5.02	10.19 | 7.79	10.19 | 7.79	10.7 | 8.12
FTS | FTS	632431 | 89824	5855287 | 6557030	0.08 | 0.09	44 | 33.03	51.61 | 39.44	51.62 | 39.43	52.95 | 40.09
FVI | FSM	192403 | 72897	1263078 | 2316467	0.43 | 0.44	3.22 | 2.39	3.74 | 2.85	3.74 | 2.86	5.55 | 4.18
G | GG	559098 | 489352	5158992 | 13485393	0.28 | 0.28	12.46 | 9.38	15.08 | 11.31	15.08 | 11.31	15.74 | 11.8
GIB.A | GIB	407242 | 80116	3157513 | 3061503	0.12 | 0.12	80.5 | 60.41	100.18 | 76.55	100.16 | 76.56	104.24 | 78.05
GIL | GIL	388478 | 127055	4230021 | 5849583	0.16 | 0.16	40.38 | 30.42	50.4 | 38.52	50.4 | 38.51	52.95 | 39.55
GOOS | GOOS	504696 | 370437	7196994 | 5321814	0.64 | 0.65	42.38 | 31.67	50.17 | 38.37	50.12 | 38.36	79.89 | 59.96
HBM | HBM	468047 | 93531	3035588 | 4150247	0.41 | 0.43	6.1 | 4.52	7.07 | 5.4	7.08 | 5.41	10.42 | 7.83
IMG | IAG	330380 | 136225	2828060 | 7338150	0.59 | 0.59	3.08 | 2.28	4.38 | 3.35	4.38 | 3.35	5.24 | 3.96
JE | JE	126472 | 26532	725152 | 664006	0.41 | 0.42	4.16 | 3.1	5.57 | 4.26	5.59 | 4.28	5.76 | 4.34
K | KGC	459455 | 204849	3423667 | 5137030	0.36 | 0.37	4.04 | 3.01	5.09 | 3.89	5.1 | 3.89	5.28 | 4
KL | KL	809303 | 265465	5883847 | 8194945	0.38 | 0.38	32.75 | 24.78	55.72 | 42.57	55.67 | 42.54	57.99 | 44.04
LAC | LAC	79010 | 19548	554294 | 409145	0.55 | 0.56	3.98 | 3	5.17 | 3.95	5.19 | 3.96	6.43 | 4.89
MFC | MFC	1055980 | 146708	9309901 | 20560263	0.18 | 0.2	19.65 | 14.73	23.83 | 18.2	23.82 | 18.2	25.18 | 18.7
MG | MGA	707903 | 225565	8590471 | 8690419	0.25 | 0.27	57.34 | 42.51	65.24 | 49.85	65.29 | 49.87	76.11 | 56.92
NEXA | NEXA	1965 | 33607	876459 | 421816	0.41 | 0.37	11 | 8.24	12.85 | 9.76	12.93 | 9.74	17.05 | 12.77
NOA | NOA	65976 | 27325	698430 | 698192	0.33 | 0.35	11.99 | 8.98	14.13 | 10.79	14.14 | 10.8	18.37 | 13.63
OR | OR	301144 | 108305	3337488 | 3325920	0.33 | 0.33	11.29 | 8.51	13.55 | 10.35	13.56 | 10.36	16.08 | 12.08
OSB | OSB	298421 | 48770	2157883 | 2327255	0.37 | 0.37	26.31 | 19.46	32.33 | 24.7	32.34 | 24.71	39.96 | 30.45

Table 12 continued from previous page

TSX Ticker | NYSE Ticker	# Trades	# Quotes	Volatility	Min. Price	Mean Price	Median Price	Max. Price
PD | PDS	275432 | 65425	1716199 | 7079878	0.58 | 0.6	2.2 | 1.65	2.43 | 1.86	2.43 | 1.87	4.05 | 3
PPL | PBA	725753 | 164168	8752195 | 16892003	0.14 | 0.15	41.9 | 31.46	48.38 | 36.95	48.32 | 36.92	50.65 | 37.93
PVG | PVG	357774 | 145260	4959405 | 12922001	0.48 | 0.49	8.85 | 6.65	13.1 | 10.01	13.11 | 10.02	13.69 | 10.4
QSR | QSR	470727 | 255245	6007156 | 5078923	0.22 | 0.21	71.83 | 53.84	90.68 | 69.29	90.67 | 69.28	93.28 | 70.46
RBA | RBA	110642 | 79826	2151942 | 2558254	0.18 | 0.18	42.64 | 31.87	43.74 | 33.41	43.74 | 33.41	49.85 | 37.90
RCI.B | RCI	705107 | 191170	5570214 | 5263925	0.16 | 0.16	65.4 | 48.68	70.04 | 53.54	70.05 | 53.54	73.82 | 55.91
RFP | RFP	8157 | 52024	1004058 | 1093796	0.44 | 0.44	8.04 | 6.03	9.13 | 6.99	9.25 | 7.08	12.80 | 9.66
SEA | SA	65739 | 57074	1567849 | 1442803	0.41 | 0.41	14.74 | 10.95	17.66 | 13.50	17.66 | 13.5	20.10 | 15.24
SHOP | SHOP	318277 | 405448	5354211 | 4998081	0.37 | 0.38	185.2 | 138.74	389.59 | 297.37	389.96 | 297.74	446.40 | 338.91
SJR.B | SJR	472788 | 111248	4124435 | 6200442	0.13 | 0.14	25.29 | 18.97	26.61 | 20.33	26.61 | 20.33	28.10 | 21.07
SLF | SLF	793187 | 175893	7900086 | 14513600	0.16 | 0.17	44.74 | 33.51	54.11 | 41.34	54.1 | 41.34	55.97 | 41.76
STN | STN	127841 | 4530	774649 | 689458	0.14 | 0.16	29.97 | 22.51	31.35 | 23.95	31.36 | 23.96	33.68 | 25.12
SU | SU	1535134 | 353457	17683883 | 57905722	0.20 | 0.21	38.64 | 28.96	40.79 | 31.17	40.8 | 31.17	46.50 | 34.86
T | TU	619142 | 111636	5611749 | 7265649	0.09 | 0.11	44.51 | 33.37	48.33 | 36.92	48.33 | 36.94	51.22 | 38.28
TA | TAC	290829 | 34737	1611785 | 1531971	0.30 | 0.30	5.78 | 4.35	8.40 | 6.43	8.41 | 6.44	10.14 | 7.61
TD | TD	1479643 | 216189	14119272 | 36534277	0.11 | 0.13	67.33 | 50.57	76.34 | 58.33	76.35 | 58.33	77.58 | 58.86
TECK.B | TECK	928365 | 326145	12283636 | 19162216	0.32 | 0.34	26.15 | 19.41	29.97 | 22.9	29.96 | 22.90	34.31 | 25.74
THO | TAHO	54648 | 34374	402457 | 1300237	0.30 | 0.32	4.54 | 3.43	4.94 | 3.75	4.94 | 3.75	5.18 | 3.93
TRI | TRI	402214 | 143245	4255310 | 5270140	0.14 | 0.14	62.92 | 47.15	84.11 | 64.25	84.1 | 64.25	88.97 | 67.26
TRP | TRP	1107841 | 245069	9958583 | 20378966	0.11 | 0.12	51.23 | 38.4	64.64 | 49.38	64.63 | 49.39	66.93 | 50.46
TRQ | TRQ	235612 | 91315	1223615 | 2755322	0.51 | 0.52	1.51 | 1.12	1.61 | 1.23	1.61 | 1.24	2.84 | 2.17
UFS | UFS	52684 | 144294	4057849 | 2763647	0.31 | 0.31	48.04 | 36.11	57.54 | 44.46	57.54 | 44.46	70.88 | 53.89
VET | VET	604493 | 110319	4984108 | 5249017	0.28 | 0.30	26.54 | 19.79	28.29 | 21.61	28.28 | 21.61	36.83 | 27.48
WEED | CGC	1828075 | 606702	13002523 | 8134021	0.54 | 0.56	37.25 | 28.01	52.97 | 40.44	52.84 | 40.37	70.98 | 52.73
WPM | WPM	622933 | 258698	9502407 | 20797947	0.25 | 0.26	24.75 | 18.55	31.38 | 23.98	31.37 | 23.98	33.85 | 25.24
YRI | AUY	337374 | 159634	2226240 | 4512850	0.43 | 0.44	2.41 | 1.79	3.33 | 2.54	3.33 | 2.55	3.78 | 2.88

Table 13. Number of days where the Augmented Dickey-Fuller test for non-stationarity is rejected at $p = 1\%$ for $\{\Gamma_t^{Short}\}$ and $\{\Gamma_t^{Long}\}$. The test was applied on each daily time series of the processes between 9:32 a.m. and 4:00 p.m. ET.

TSX Ticker | NYSE Ticker	Short	Long	TSX Ticker | NYSE Ticker	Short	Long
ABX | GOLD	0	0	IMG | IAG	0	0
AEM | AEM	0	0	JE | JE	1	0
AGI | AGI	0	0	K | KGC	0	0
AQN | AQN	0	0	KL | KL	0	0
ATP | AT	3	5	LAC | LAC	0	2
BAM.A | BAM	0	0	MFC | MFC	0	0
BB | BB	0	0	MG | MGA	0	0
BCB | COT	0	0	NEXA | NEXA	4	2
BCE | BCE	0	0	NOA | NOA	1	0
BMO | BMO	0	0	OR | OR	0	0
BNS | BNS	0	0	OSB | OSB	0	0
BTE | BTE	0	0	PD | PDS	0	0
BXE | BXE	6	12	PPL | PBA	0	0
CAE | CAE	0	0	PVG | PVG	0	0
CCO | CCJ	0	0	QSR | QSR	0	0
CLS | CLS	0	0	RBA | RBA	0	0
CM | CM	0	0	RCI.B | RCI	0	0
CNQ | CNQ	0	0	RFP | RFP	0	0
CNR | CNI	0	0	SEA | SA	0	0
CNU | CEO	4	19	SHOP | SHOP	0	0
CP | CP	0	0	SJR.B | SJR	0	0
CPG | CPG	0	0	SLF | SLF	0	0
CVE | CVE	0	0	STN | STN	1	2
ECA | ECA	0	0	SU | SU	0	0
EDR | EXK	0	0	T | TU	0	0
ELD | EGO	0	0	TA | TAC	0	0
ENB | ENB	0	0	TD | TD	0	0
ERF | ERF	0	0	TECK.B | TECK	0	0

Table 13 continued from previous page

TSX Ticker | NYSE Ticker	Short	Long	TSX Ticker | NYSE Ticker	Short	Long
FNV | FNV	0	0	THO | TAHO	0	0
FR | AG	0	0	TRI | TRI	0	0
FTS | FTS	0	0	TRP | TRP	0	0
FVI | FSM	0	0	TRQ | TRQ	1	1
G | GG	0	0	UFS | UFS	0	0
GIB.A | GIB	0	0	VET | VET	0	0
GIL | GIL	0	0	WEED | CGC	0	0
GOOS | GOOS	0	0	WPM | WPM	0	0
HBM | HBM	0	0	YRI | AUY	0	0

Note: Since we observe a low number of days where $\{\Gamma_t^{Short}\}$ and $\{\Gamma_t^{Long}\}$ are not stationary, the mean-reversion risk is minimal in our strategy for almost all pairs. Even though that risk is very low, our risk management strategy still implements circuit breakers with a timer and a stop-loss to capture as much arbitrage opportunities as possible.

Table 14. Variable definitions and symbols used in the regression analysis.

Variable name	Symbol	Definition*
Daily average net profits per selected pair (C$)	$\overline{profits}_{k,d}$	$\dfrac{\sum_{n=1}^{N} profits_d^{(n)}}{\lvert P_d \rvert}$, where $profits_d^{(n)}$ are the net profits in C$ generated by the pair of cross-listed stocks n on day d, and $\lvert P_d \rvert$ is the cardinality of the set of selected pairs on that day from our machine learning methodology. Non-selected pairs have $profits_d^{(\cdot)} = 0$.
Intraday mid-price volatility	$vol_{k,d}$	$\dfrac{\sum_{n=1}^{N} w_{k,d}^{(n)} vol_{k,d}^{(n)}}{\sum_{n=1}^{N} w_{k,d}^{(n)}}$, where $w_{k,d}^{(n)} = \sigma_{k,d}^{(n)}/\mu_{k,d}^{(n)}$ and $\sigma_{k,d}^{(n)}$ and $\mu_{k,d}^{(n)}$ are the respective standard deviation and mean of the mid-price series of stock n at exchange k and day d.
Average intraday mid-price volatility of stocks	$\overline{vol}_{stocks,d}$	$\dfrac{vol_{TSX,d} + vol_{NYSE,d}}{2}$
Bid-ask spread	$spread_{k,d}$	$\dfrac{\sum_{n=1}^{N} w_{k,d}^{(n)} spread_{k,d}^{(n)}}{\sum_{n=1}^{N} w_{k,d}^{(n)}}$, where $spread_{k,d}^{(n)}$ is the mean bid-ask spread series (in bps) of stock n, at exchange k and day d.
Average bid-ask spread for stocks	$\overline{spread}_{stocks,d}$	$\dfrac{spread_{TSX,d} + spread_{NYSE,d}}{2}$
Trading volume	$trade_{k,d}$	$\dfrac{\sum_{n=1}^{N} w_{k,d}^{(n)} trade_{k,d}^{(n)}}{\sum_{n=1}^{N} w_{k,d}^{(n)}}$, where $trade_{k,d}^{(n)}$ is the trading volume of stock n, at exchange k and day d.
Average trading volume for stocks	$\overline{trade}_{stocks,d}$	$\dfrac{trade_{TSX,d} + trade_{NYSE,d}}{2}$
Number of L1 messages	$messages_{k,d}$	$\dfrac{\sum_{n=1}^{N} w_{k,d}^{(n)} messages_{k,d}^{(n)}}{\sum_{n=1}^{N} w_{k,d}^{(n)}}$, where $messages_{k,d}^{(n)}$ is the number of L1 messages of stock n, at exchange k and day d.
Average number of L1 messages for stocks	$\overline{messages}_{stocks,d}$	$\dfrac{messages_{TSX,d} + messages_{NYSE,d}}{2}$

*A cross-listed stock n listed at exchange k has a daily traded value on day d of $w_{k,d}^{(n)} = \sum_{q=1}^{Q_{k,d}^{(n)}} \nu_{k,q}^{(n)} S_{k,d}^{(n)}$ for $Q_{k,d}^{(n)}$ the number of trades that resulted from the limit order strategy. $\nu_{k,q}^{(n)}$ is the volume of the q^{th} trade and $S_{k,d}^{(n)}$ is the stock's value when the trade ended. Note that $Q_{k,d}^{(\cdot)} = 0$ for every non-selected pair on day d.

Table 15. Pearson correlation matrix of the variables used in the regression analysis. Bold values are statistically different from 0.

	$\overline{profits}$	vol_{TSX}	vol_{NYSE}	vol_{CME}	$spread_{TSX}$	$spread_{NYSE}$	$spread_{CME}$	$trade_{TSX}$	$trade_{NYSE}$	$trade_{CME}$	$messages_{TSX}$	$messages_{NYSE}$	$messages_{CME}$
$\overline{profits}$	**1.000**	**0.474**	**0.490**	0.134	**0.613**	**0.669**	-0.114	0.031	**0.228**	0.118	**0.344**	**0.228**	**0.394**
vol_{TSX}		**1.000**	**0.983**	**0.290**	**0.462**	**0.471**	0.007	**0.242**	**0.607**	0.025	**0.354**	**0.249**	0.150
vol_{NYSE}			**1.000**	**0.322**	**0.432**	**0.447**	0.010	**0.263**	**0.609**	0.019	**0.390**	**0.287**	0.180
vol_{CME}				**1.000**	0.117	0.140	**0.244**	0.021	0.133	0.167	0.157	0.125	**0.258**
$spread_{TSX}$					**1.000**	**0.932**	-0.137	**-0.315**	0.160	0.126	-0.197	**-0.304**	0.180
$spread_{NYSE}$						**1.000**	-0.092	-0.165	0.191	0.117	-0.097	**-0.190**	**0.237**
$spread_{CME}$							**1.000**	0.017	-0.052	-0.090	0.066	0.101	**-0.351**
$trade_{TSX}$								**1.000**	**0.742**	-0.054	**0.702**	**0.688**	-0.033
$trade_{NYSE}$									**1.000**	0.026	**0.630**	**0.516**	-0.005
$trade_{CME}$										**1.000**	-0.014	-0.060	**0.209**
$messages_{TSX}$											**1.000**	**0.845**	0.170
$messages_{NYSE}$												**1.000**	**0.192**
$messages_{CME}$													**1.000**

Appendix B. Practical considerations for strategy implementation

The equilibrium value of the relative spread, τ^i, $i \in \{Short, Long\}$ can be computed *a posteriori* at the end of the day. However, in practice, these quantities need to be known in real time to find the arbitrage opportunities. To account for overnight basis adjustment, a simple approximation can be the sample average of the $\{\Gamma_t^i\}$ processes during the first minutes of a trading day before starting the strategy. We remove the first two minutes of each trading day to let the prices converge to their daily equilibrium level.

The approximation is then used as the first value of τ^i when the strategy starts. From that starting point, the approximation is following a running average of $\{\gamma_t^i\}$ at every L1 event in one of the three exchanges for a given stock and currency futures. Note that the strategy needs a constant equilibrium value from the opening trades to the closing trades, meaning that the τ^is are not updated when positions are still open for a given pair.

The strategy assumes that the synthetic spreads return exactly to equilibrium at their respective time t'. Because of market frictions (mainly discrete stock prices), there is a null probability that the synthetic spreads would converge exactly to τ^i at any time, so bounds around equilibrium are necessary to close the positions. To solve this issue, we add another parameter $\beta \in \mathbb{R}_{>0}$ that controls when processes are near enough to their respective equilibrium to close the positions within reasonable bounds. The practical definition of t' becomes:

$$t'_{Market,Short} \equiv \underset{s>t}{\arg\min} \left\{ s \mid \gamma_s^{Short} \in [\tau^{Short} - \beta(\tau^{Short} - \kappa_s^{Under}), \tau^{Short} + \beta(\tau^{Short} - \kappa_s^{Under})] \right\}$$

for the process $\{\Gamma_t^{Short}\}$ with the market order-based strategy,

$$t'_{Market,Long} \equiv \underset{s>t}{\arg\min} \left\{ s \mid \gamma_s^{Long} \in [\tau^{Long} - \beta(\kappa_s^{Over} - \tau^{Long}), \tau^{Long} + \beta(\kappa_s^{Over} - \tau^{Long})] \right\}$$

for the process $\{\Gamma_t^{Long}\}$ with the market order-based strategy,

$$t'_{Limit,Short} \equiv \underset{s>t}{\arg\min} \left\{ s \mid \gamma_s^{Short} \in [\tau^{Short} - \beta(\tau^{Short} - \tilde{\kappa}_s^{Over}), \tau^{Short} + \beta(\tau^{Short} - \tilde{\kappa}_s^{Over})] \right\}$$

for the process $\{\Gamma_t^{Short}\}$ with the limit order-based strategy,

$$t'_{Limit,Long} \equiv \arg\min_{s>t} \left\{ s \mid \gamma_s^{Long} \in [\tau^{Long} - \beta(\tilde{\kappa}_s^{Under} - \tau^{Long}), \tau^{Long} + \beta(\tilde{\kappa}_s^{Under} - \tau^{Long})] \right\}$$

for the process $\{\Gamma_t^{Long}\}$ with the limit order-based strategy.

The smaller the β, the nearer the processes need to be to equilibrium to close the positions and the closer the practical strategy gets to the theory.

The volumes sent to the market by the strategy are round lots because of the higher costs related to sending odd lot orders, meaning that the minimum volume that can be used in the strategy is 100 stocks on both stock exchanges. To capture as much arbitrage opportunities as possible without heavily impacting the price discovery processes, we dynamically determine the orders' volume following the first level volumes available in the LOB of the exchanges for a given pair of stocks. The orders' volume sent to both markets is limited by the less active one, since for one stock in Exchange 1's market, we take a position of $1/\tau$ stocks in Exchange 2's market. We have observed that τ^i does not deviate far enough from 1 to send a different number of lots in both markets for the same arbitrage opportunity. Therefore, the implemented strategy sends the same volumes to both stock exchanges.

Defining $\tilde{\nu}_{Exchange,t}^{Side}$ as the median volume on the first LOV level on $Side \in \{Bid, Ask\}$ in $Exchange \in \{1,2\}$ based on the last 500 L1 updates preceding time t.[11] The volume sent to both markets at time t for any cross-listed stock, ν_t, is computed as either:

$$\nu_t = 100 \max\left(\min\left(\left\lfloor \frac{\tilde{\nu}_{1,t}^{Bid}}{100} \right\rfloor, \left\lfloor \frac{\tilde{\nu}_{2,t}^{Ask}}{100} \right\rfloor \right), 1 \right),$$

or

$$\nu_t = 100 \max\left(\min\left(\left\lfloor \frac{\tilde{\nu}_{1,t}^{Ask}}{100} \right\rfloor, \left\lfloor \frac{\tilde{\nu}_{2,t}^{Bid}}{100} \right\rfloor \right), 1 \right),$$

depending on whether market or limit orders are used, and whether a long or short position is opened or closed in SPDR.

[11] We tested the robustness of the strategy with respect to the median volume by using 250, 1,000 and 2,500 L1 updates. The profitability did not significantly change.

As mentioned previously, currency futures are used to hedge positions from currency risk. The optimal position in that instrument is given by equation (7) at any time during the strategy's execution. To follow that position as closely as possible, the Micro CAD/USD futures contract with a nominal of C$10,000 is employed, which is approximated by dividing the prices of the continuous futures by 10, because of its nominal of C$100,000.

Let $\hat{\nu}_{FX,t} \in \mathbb{Z}$ be the number of currency futures contracts needed at time t which best approximates the position size theoretically needed at the FX Exchange at that time, $\nu^*_{FX,t}$, without under-hedging the aggregated position in Exchange's 2 market. We compute its value as:

$$\hat{\nu}_{FX,t} = \begin{cases} \lfloor \nu^*_{FX,t} \rfloor \text{ if } \nu^*_{FX,t} \leq 0 \\ \lceil \nu^*_{FX,t} \rceil \text{ if } \nu^*_{FX,t} > 0 \end{cases}, \forall t.$$

Because of high nominal value of the futures, we cannot perfectly hedge Exchange 2's positions. In the market order strategy, only market orders are used to follow as much as possible $\hat{\nu}_{FX,t}$ during the strategy's execution. In the limit order strategy, limit orders are sent to the top-of-the book prices, or canceled, or updated at every market event modifying $\hat{\nu}_{FX,t}$ to achieve the same goal. Latency makes it more complicated to get exactly a volume of $\hat{\nu}_{FX,t}$ at all times.[12]

To mitigate the mean-reversion risk and the non-execution risk, specifically for the limit order strategy, a timer of 15 minutes is used to cancel any order and close any position resulting from opening a position in the synthetic spread (SPRD) using marketable limit orders. The timer starts when orders are sent to the markets and ends only when the orders are filled, and the positions are closed. Along the same vein, stop-losses are also implemented so that if the prices in the LOB level one diverge drastically from pending limit order prices, these orders would be canceled, and any opened position would be closed with marketable limit orders. No new positions are opened 15 minutes before market close.

[12] Trading the hedging instrument does not directly lead to significant losses or gains, but is necessary to mitigate currency risk in both strategies. Slippage of market orders for the hedging instrument is insignificant to the profitability of the strategy. Slipping does not occur with limit orders, but the non-execution risk can generate a non-optimal hedging position, even more so when latency is considered.

Even though the strategy is built to be theoretically profitable for every pair, the cross-listed stocks' characteristics could lead to unprofitable trades. To determine how the underlying factors of a profitable pair differ from the ones of a unprofitable pair, we resort to supervised learning. The resulting machine learning algorithm allows us to predict the future profitability of our pairs, thus enabling dynamic pair selection and optimizing the strategy's performance by filtering out potentially unprofitable pairs.

Specifically, we utilize a decision tree algorithm, because of its interpretability. We apply this nonparametric model to predict if a given pair will be profitable in the next period based on the data in previous periods. We treat this problem as a dynamic binary classification task where the output of the model at each period is either profitable or unprofitable for each pair in the universe Ω during the next period. See Appendix C for more details on the pair selection method using the decision tree algorithm.

Appendix C. Decision tree learning for recurrent pair selection

For D the number of days in the data, $\mathcal{D} = \{1,2,\dots,D\}$ the daily indices, and $\ell \in \mathbb{Z}_{<D/3}$ the period length at which the pairs are recurrently selected in Ω. Pair selection is done every ℓ days throughout the data, beginning after two periods, since the first two periods are needed for the first decision tree to be trained. Set $M = \left\lfloor \frac{D}{\ell} \right\rfloor$ with the model training indices $\mathcal{M} = \{2\ell, 3\ell, \dots, M\ell\}$. Define $\{\mathbf{X}_{p,d}\}_{t\in\mathcal{D}}$, $\mathbf{X}_{p,d} \in \boldsymbol{\mathcal{X}} \subseteq \mathbb{R}^w$ the multivariate stochastic process of the w daily predictive features with time series $\{\mathbf{x}_{p,d}\}_{d\in\mathcal{D}}$, and $\{\Pi_{p,d}\}_{d\in\mathcal{D}}$, $\Pi_{p,d} \in \mathbb{R}$ the net daily profit process of pair $p \in \{1,2,\dots,n\}$ with time series $\{\pi_{p,d}\}_{d\in\mathcal{D}}$ generated by the strategy on pair p during day d.

Let's also define $\{Y_{p,m}\}_{m\in\mathcal{M}}$, $Y_{p,m} \in \mathcal{Y} \equiv \{-1, 1\}$, the profitability class process of pair p, with time series $\{y_{p,m}\}_{m\in\mathcal{M}}$ where $y_{p,m} = -1$ when the pair p is unprofitable and $y_{p,m} = 1$

when it is profitable during $\{t \mid m-\ell+1 \leq d \leq m\}$. The time series is computed as follows:

$$y_{p,m} = \begin{cases} -1 & \text{if } \sum_{t=0}^{\ell-1} \pi_{p,m-t} \leq 0 \\ 1 & \text{if } \sum_{t=0}^{\ell-1} \pi_{p,m-t} > 0 \end{cases}, \forall p, m.$$

The decision tree's goal is to learn a series of $M-1$ functions $H_m : \boldsymbol{\mathcal{X}} \mapsto \mathcal{Y}$, where each one maps the features of all the pairs in $\{t \mid m-2\ell+1 \leq d \leq m-k\}$ to their respective profitability class during the next period $\{t \mid m-\ell+1 \leq d \leq m\}$ for all $m \in \mathcal{M}$. To do so, we use the arithmetic mean of the daily features' time series as inputs to the model. Hence, the set of training tuples for the decision tree at time m is given by:

$$\mathcal{S}_m = \left\{ \left(\frac{\sum_{t=0}^{\ell-1} \mathbf{x}_{1,m-\ell-t}}{\ell}, y_{1,m} \right), \left(\frac{\sum_{t=0}^{\ell-1} \mathbf{x}_{2,m-\ell-t}}{\ell}, y_{2,m} \right), \ldots, \left(\frac{\sum_{t=0}^{\ell-1} \mathbf{x}_{n,m-\ell-t}}{\ell}, y_{n,m} \right) \right\}, \forall m$$

Based on a set $\mathcal{S}_m$, the decision tree tries to find H_m using Gini's impurity criterion and information gain. To avoid overfitting issues, we limit the algorithm to a maximum depth of two. The resulting classification function, $\widehat{H}_m$, is then used to predict the profitability of each pair in the next interval $\{t \mid m-\ell+1 \leq d \leq m\}$ from the most recent features:

$$\widehat{H}_m \left(\frac{\sum_{t=0}^{\ell-1} \mathbf{x}_{p,m-t}}{\ell} \right) = \widehat{y}_{p,m+\ell}, \ \forall p,m.$$

Hence, we can select the set of pairs that will be traded throughout the next interval, which is given by:

$$P_m = \{j \mid \widehat{y}_{j,m+\ell} = 1, 1 \leq j \leq n\}, \ \forall m.$$

The decision tree is completely retrained only on the corresponding training examples at each $m \in \mathcal{M}$, so that only local patterns are used in the prediction of the pairs' profitability. Note that we cannot launch the pair selection method until $d = 2\ell$, because the first ℓ days are used to generate the features, and the following ℓ days are used to compute the profitability of the pairs. Together, these features and profitability values from the first set of training examples on which we train the first decision tree at $d = 2\ell$.

The daily features that compose $\{\mathbf{X}_{p,d}\}_{d \in \mathcal{D}}$ are:

- bid-ask spread,
- total trading volume,
- ratio of the number of trades per quote,
- mid-price's coefficient of variation,
- total number of trades and quotes, and
- a measure of the previous period's profitability,

with $\ell = 3$ days. The profitability prediction accuracy of the decision tree at time $m \in \mathcal{M}$, defined as A_m, is computed as follows:

$$A_m = \frac{\sum_{p=1}^{n} \mathbb{I}_{\{\widehat{y}_{p,m+\ell} = y_{p,m+\ell}\}}}{n}, \ \forall m,$$

where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. The accuracy at each period is presented in the next figure.

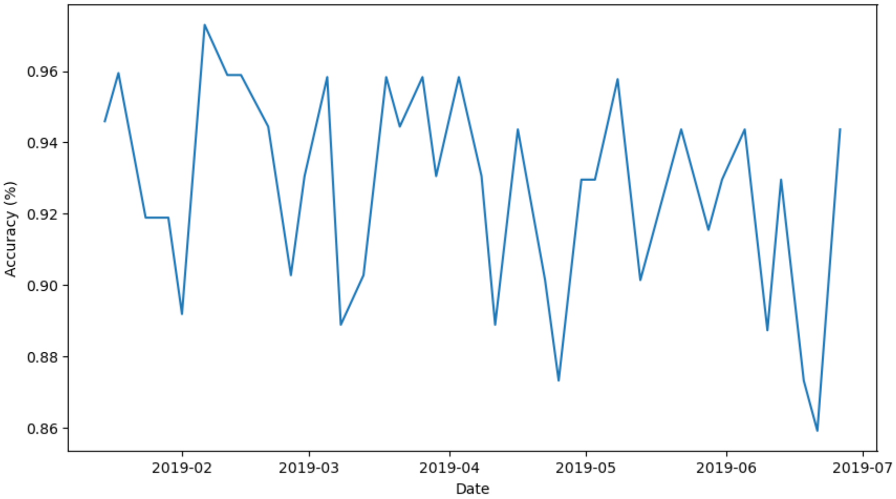

Fig. 9. Prediction accuracy A_m, $\forall m \in \mathcal{M}$ of the dynamic decision tree approach from January 15 to June 20, 2019, computed every $\ell = 3$ days for every pair in Ω.

From Figure 9, it can be observed that the methodology predicts that the next three days of each pair will be profitable at an average of 92% accuracy. Also, the predicted accuracy does not vary very much in the period of analysis. This process is repeated until the end of our data. Figure 10 presents the selected pairs in time for our portfolio. It can be observed that only 36 pairs (in green) were selected at least one time.

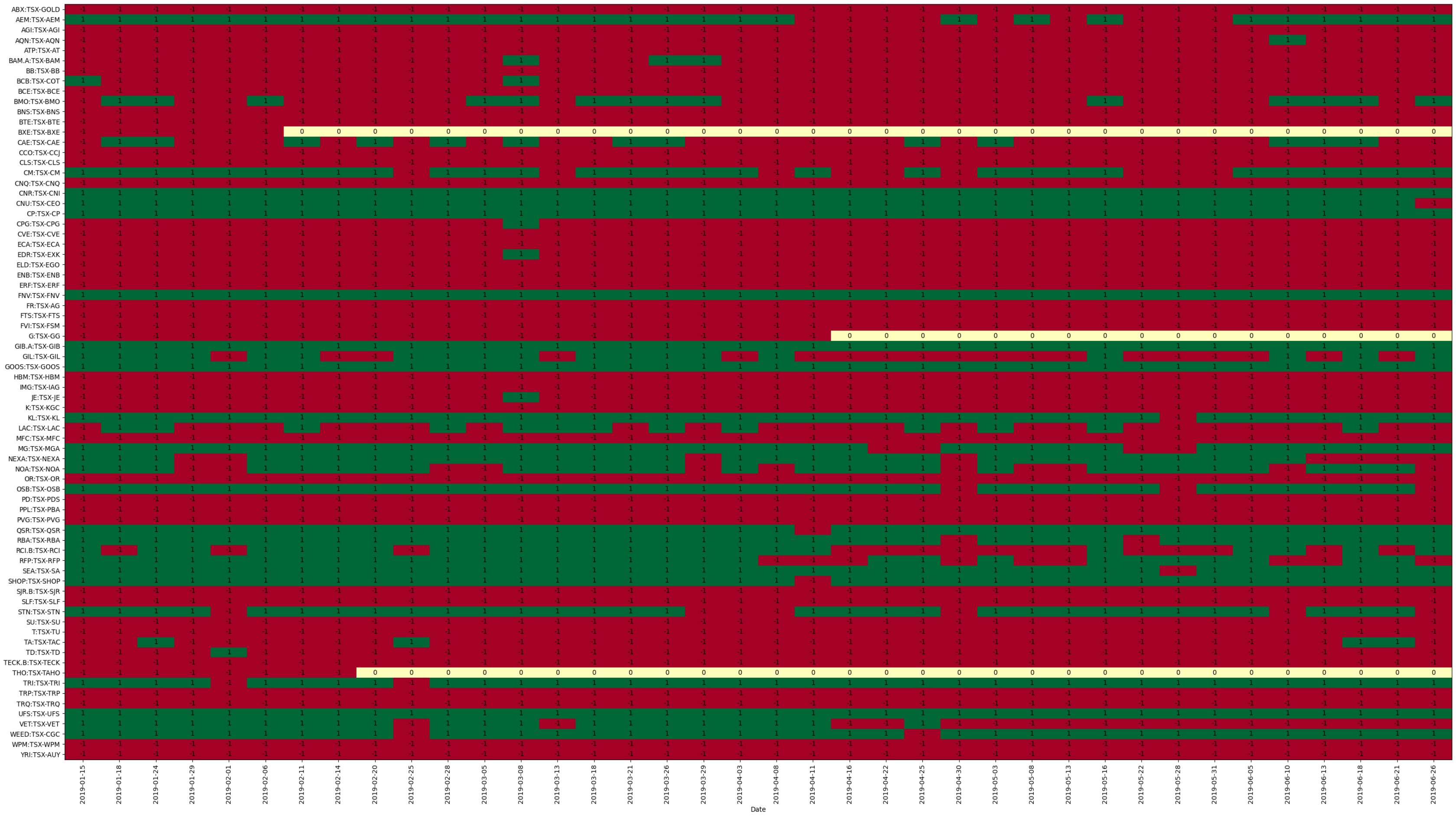

Fig. 10. Predicted profitability of each pairs in time generated by the dynamic decision tree–based approach from January 15 to June 28, 2019: $\widehat{y}_{p,m_\ell}$, $\forall p \in \{1,2,\ldots,74\}$, $\forall m \in \{6,9,\ldots,114\}$. The selected pairs in time, P_m, are in green, non-selected pairs are in red, and pairs where at least one stock is delisted are in yellow. Overall, 36 pairs were selected by the model at least one time.

Figures 11, 12, 13 represent the decision trees learned for pair selection at the beginning of our period of analysis on 2019-01-15, in the middle period on 2019-03-29, and towards the end on 2019-06-13, respectively. Each rectangle in a tree is a node with the best rule minimizing the Gini impurity of the corresponding child nodes. A rule is a criterion that splits the feature space into distinct subspaces. Feature vectors that fall within one of the resulting subspaces are then passed to the corresponding child node. Feature vectors that respect the interval specified by the rule in a node continue to the bottom left, and if they do not, they continue to the right until they arrive to a leaf where the prediction takes place. The learned rules are the first line of each non-leaf node. Leaves have no rules and are located at the bottom of the trees. The prediction made at the leaves is the most predominant class in the node's sub data set, where the number of instances of each class is given by the vector "value."

To determine the pair selection variables, well-established stylized facts are dynamically fed *ex ante* into a decision tree using three days of high-frequency data. The information set is restricted to variables from tick trades and limit order level one, and the target is the strategy's daily profitability class for each pair. The tree learning is done after markets close and is used during the three following intraday activities. Most of the time, two conceptually appealing stylized facts drive the pair selections: the bid-ask spread, an important component of endogenous liquidity providers profitability Ait-Sahalia and Saglam [1]; Brogaard et al. [9], and the number of messages, tightly linked to liquidity Hasbrouck and Saar [36]; Hendershott et al. [37]. The three decision trees below exemplify a recurring decision tree structure. The pair selection methodology based on them generates more than satisfactory results, given their out-of-sample high profitability prediction accuracy and the excellent stability in the performance through time (see Figure 9). This confirms that the features selected by the decision trees are reliable predictors of the profitability of each pair traded.

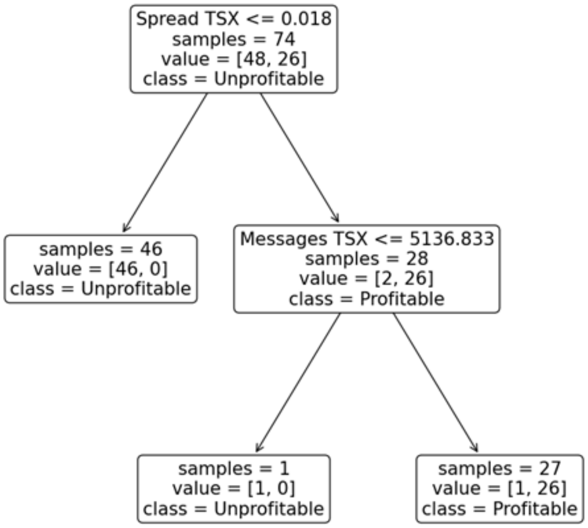

Fig. 11. Decision tree for pair selection on 2019-01-15

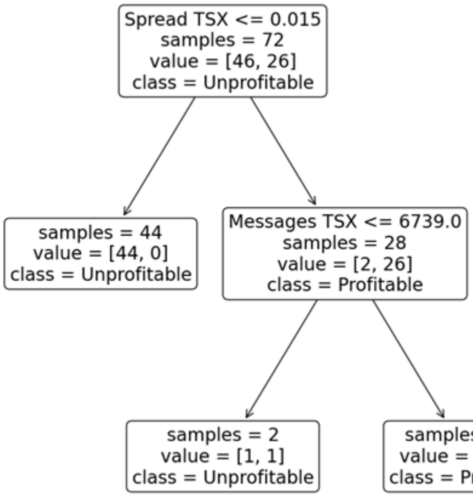

Fig. 12. Decision tree for pair selection on 2019-03-29

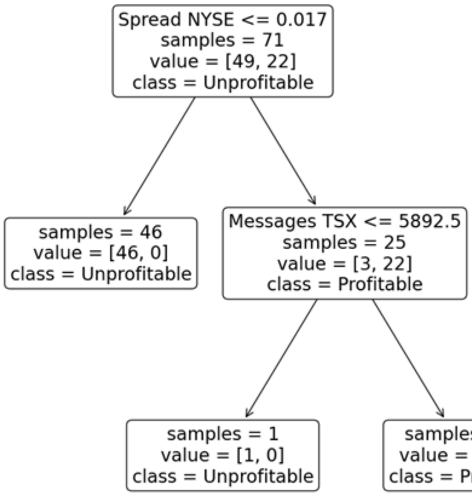

Fig. 13. Decision tree for pair selection on 2019-06-13

Appendix D. Study of Wah's strategy

We repeat the experiments done in Section 6 for Budish et al. [12], but with the strategy of Wah [60]. Once again, it is implemented with the observed theoretical settings and minor modifications to adapt it to our data. Prices at NYSE are continuously transferred to C$ following the CAD/USD futures observed at CME. In addition, we used two hypotheses employed in the model: There is an absence of latency, and opened positions at an exchange can be immediately closed at another exchange, resulting in a trade. There is a small nuance in the case of Wah [60]: MIDAS data is recorded at a single point of observation, meaning that the effect of latency on information observation is already considered. There still remains the latency of the orders. Table 16 presents the results obtained on our data with the strategy of Wah [60]. The second column of Table 16 presents the results that are obtained following as closely as possible the respective theoretical framework that cannot be replicated in practice. In the next two columns, latency is considered.

Wah [60] utilizes direct-feed data from MIDAS, a platform at the U.S. Securities and Exchange Commission (SEC) that provides access to order and quote messages on all U.S. stock exchanges. Cross-market arbitrage opportunities are analyzed from 11 U.S. equities exchanges. The author assumes there is a single infinitely fast latency arbitrageur. When the arbitrageur detects a latency arbitrage opportunity, the strategy is to submit market orders to the exchanges involved in the cross-market arbitrage opportunity. The data used by Wah [60] includes market orders for the 495 tickers of the S&P 500 from January 1, 2014 to December 31, 2014. Latency arbitrage opportunities across these exchanges were observed to happen very frequently during that year and they generated a profit exceeding US$3.03 billion to the infinitely fast latency arbitrageur.[13]

When we look at column 2 of Table 16, the results of the original model with our data generate a gross profit of C$4.7 million for 74 stocks in two exchanges for six months.[14] If we

[13] 46 tickers from the Russell 2000 were also studied but their profits are not included in the US$3.03 billion result.

[14] In this section we do not use the futures contracts for hedging the exchange rate. We do however use the exchange rate updates continuously to obtain pure variations in stock prices between Toronto and NY exchanges.

Table 16. Wah [60] strategy on our data

1	**2**	**3**	**4**
Model	**Wah Original w/o Latency**	**Wah Original w/ 1x Latency**	**Wah Original w/ 3x Latency**
Latency multiplier	0 *	1	3
Pair selection	No	No	No
Gross profit	$4,677,764.64	$4,625,043.07	$4,305,331.72
Loss	$0.00	-$282,933.09	-$308,364.36
Trading fees	-$2,674,499.10	-$2,832,777.28	-$2,721,234.00
Trading rebates	$0.00	$0.00	$0.00
Total net profit	$2,003,265.54	$1,509,332.70	$1,275,733.36
Mean daily net profit	$17,572.50	$13,239.76	$11,190.64
Median daily net profit	$17,083.02	$12,782.75	$10,672.95
Mean daily net profit per pair, per day	$237.47	$178.92	$151.22
$p - value$ Kolmogorov-Smirnov test		1.00	1.00
1^{st} most profitable day (date - profit)	2019/01/17 $35,222.68	2019/01/17 $30,985.29	2019/01/17 $29,793.92
5^{th} most profitable day (date - profit)	2019/01/30 $27,788.04	2019/01/30 $23,643.17	2019/05/07 $19,803.40
1^{st} most unprofitable day (date - profit)	2019/04/11 $6,013.47	2019/06/17 $1,470.35	2019/04/11 $502.20
5^{th} most unprofitable day (date - profit)	2019/04/05 $7,843.25	2019/04/15 $4,424.84	2019/06/20 $3,535.60
Average time in trade	00:00.0	00:00.0	00:00.0
# Net profitable trades	158,647	154,718	155,543
# Net unprofitable trades	49,703	76,095	79,131
# Trades	208,350	230,813	234,674
% Net profitable trades	76.14%	67.03%	66.28%
Average volume per trade	2366.07	2,380.48	2,249.33
Average net profit per trade	$9.61	$6.54	$5.44
Average profit per net profitable trade	$15.71	$15.96	$14.89
Average profit per net unprofitable trade	-$9.83	-$12.61	-$13.14

extend these results to eleven exchanges with 495 stocks over one year, this generates about C$0.76 billion (US$0.58 billion) in the year 2019.[15]

[15] ($4,677,764.64 × (495/74) × 11 × (252/114) = $760,852,059.40).

The main difference with Wah's original study can be explained by the characteristics of the stocks in the two studies and by the relative sizes of the exchanges. To have a comparable market environment to Wah [60], when generating the C$0.76 billion result we assumed there is only one very fast arbitrageur in colocation in only one exchange and trading in the 11 exchanges. If we extend the possibility that the trading activities are generated by the very fast arbitrageur in colocation in the eleven markets, we obtain about C$3.8 billion (US$2.9) in annual gross profits ($0.76×5)[16], which is fairly close to the US$3 billion reported in the paper. We also observe, in column 2 of Table 16, that the trading costs represent more than half of the gross profit generated by the strategy. Only around 76% of the arbitrage opportunities cover the trading costs. In that sense, Budish et al. [12] approach allows to better select arbitrage opportunities. This is also true when comparing average net profits per trades between these two strategies.

Assuming an infinitely fast arbitrageur cannot correspond to any known trading application in the real world. In the next columns, latency is incorporated in the trading environment. This results in a statistically significant (see $p-values$) decrease of 25% in net profitability. The drop in profitability is even greater at 36% when latency is tripled from the based value. Once again, this demonstrates the importance of latency in HFT profitability. Ignoring this practical aspect inflates the reported profits.

As in Budish et al. [12], Wah [60] considers that a trade occurs when two opposite positions are taken in different exchanges. We abandon this hypothesis, meaning that an opposite position at the same exchange has to be taken in order to lead to a trade. The results generated by this last strategy in this more practical market environment are presented in Table 17.

The outcomes obtained by the strategy of Wah [60] in Table 17 lead to the same observations that were previously made based on Budish et al. [12] results in Section 6: The strategy does not generate any net profit. Profits statistically decrease whenever latency is

[16] Here we assume the arbitrageur exploits 55 links between the exchanges even if she receives information at one single observation point. A better approximation should consider the real volumes of arbitrage between the exchanges.

Table 17. Wah [60] strategy on our data with practical hypotheses

1 **Model**	**2** **Wah Practical w/o Latency**	**3** **Wah Practical w/ 1x Latency**	**4** **Wah Practical w/ 3x Latency**
Latency multiplier	0*	1	3
Pair selection	No	No	No
Gross profit	\$343,498.69	\$337,486.65	\$380,996.00
Loss	-\$350,969.16	-\$346,625.18	-\$393,820.69
Trading fees	-\$5,906.46	-\$5,811.90	-\$6,877.05
Trading rebates	\$0.00	\$0.00	\$0.00
Total net profit	-\$13,376.93	-\$14,950.43	-\$19,701.74
Mean daily net profit	-\$117.34	-\$131.14	-\$172.82
Median daily net profit	-\$30.39	-\$29.40	-\$40.99
Mean daily net profit per pair, per day	-\$1.59	-\$1.77	-\$2.34
$p - value$ Kolmogorov-Smirnov		0.97	0.99
1st most profitable day (date - profit)	2019/06/28 \$2,447.42	2019/06/28 \$2,673.73	2019/06/28 \$2,462.57
5th most profitable day (date - profit)	2019/06/25 \$243.73	2019/06/25 \$177.66	2019/06/25 \$273.35
1st most unprofitable day (date - profit)	2019/05/15 -\$4,254.55	2019/05/15 -\$4,728.90	2019/05/16 -\$2,931.57
5th most unprofitable day (date - profit)	2019/04/05 -\$728.58	2019/06/23 -\$922.47	2019/06/03 -\$1,350.77
Average time in trade	118.17:19:47	119.05:04:35	122.17:30:38
# Net profitable trades	513	498	728
# Net unprofitable trades	527	512	756
# Trades	1,040	1,010	1,484
% Net profitable trades	49.33%	49.31%	49.06%
Average volume per trade	549.21	556.50	448.69
Average net profit per trade	-\$12.86	-\$14.80	-\$13.28
Average profit per net profitable trades	\$665.73	\$673.80	\$520.00
Average profit per net unprofitable trades	-\$673.43	-\$684.58	-\$526.80
Total Short Inventory Remaining	\$4,705,786,414.13	\$4,693,771,499.50	\$4,643,810,959.12
Total Long Inventory Remaining	\$3,587,847,145.39	\$3,578,678,416.90	\$3,540,608,627.73

introduced in the testing environment. A great inventory has also been accumulated during the six months, even more so than the strategy of Budish et al. [12], for the same reasons. Overall, by not considering practical trading aspects such as latency or real market functioning, Wah [60] inflated latency arbitrage's profitability.

Appendix E. Execution rules

1. Each limit order has a standing quantity that must be executed before the order is executed.
2. That standing quantity is computed from the following steps:
   a. If the limit order's price of a buy/sell order is equal to the best bid/ask price, the order's standing quantity becomes the current best bid/ask volume.
   b. If the limit price of a buy/sell order is below/above the best bid/ask price, the order's standing quantity is undefined. In that instance, the trading and quoting emulator waits for the limit order's price to be equal to the best bid/ask price and it sets the standing quantity according to 2.a.
   c. If the limit order's buy/sell price is above/below the best bid/ask price, the order is filled.
   d. If the standing quantity has been defined for a limit order, it can only be changed by a future execution.
3. A limit order can be executed by a trade occurring at the limit order's price. The standing quantity must be executed first. If it has been executed completely, then the limit order can be executed. If the remaining trade size is not large enough to fill the limit order's size, then a partial filling occurs. Limit orders with an undefined standing quantity cannot be executed by a trade.
4. A limit order can be executed when the best ask/bid price becomes lower/greater than the buy/sell limit order's price. This also holds for limit orders with undefined standing quantities.

5. A limit order is filled when the best bid/ask price becomes lower/greater than the buy/sell limit order's price, regardless of its standing quantity. This also holds for limit orders with undefined standing quantities.

The trading and quoting emulator is conservative in some regards, especially considering the static standing quantity that must be executed before the corresponding limit order, because it ignores cancellations decreasing that quantity after the order has been placed, which follows from rules 1 and 2.a. Also, whenever a limit order is placed deeper than LOB level 1 and its price becomes the top of the book after some time, the limit order is put at the end of the queue of all the orders also at the new level 1 regardless of its actual position in that queue, which follows from rules 2.a and 2.b.

Second Article.

The Profitability of Lead-Lag Arbitrage at High-Frequency

by

Cédric Poutré[1], Georges Dionne[2], and Gabriel Yergeau[2]

([1]) Department of Mathematics and Statistics
Université de Montréal

([2]) Department of Finance
HEC Montréal

This article was submitted to International Journal of Forecasting and passed the first review round.

The main contributions of Cédric Poutré for this article are presented.

- Original idea behind this article;
- Creation of the theoretical models;
- Design of the trading strategies;
- Production of numerical results;
- Writing of all the manuscript.

Georges Dionne helped with the manuscript.

Gabriel Yergeau also helped with the manuscript and the data integration.

Abstract. Any lead-lag effect in an asset pair implies the future returns on the lagging asset have the potential to be predicted from past and present prices of the leader, thus creating statistical arbitrage opportunities. We utilize robust lead-lag indicators to uncover the origin of price discovery and we propose an econometric model exploiting that effect with level 1 data of limit order books (LOB). We also develop a high-frequency trading strategy based on the model predictions to capture arbitrage opportunities. The framework is then evaluated on six months of DAX 30 cross-listed stocks' LOB data obtained from three European exchanges in 2013: Xetra, Chi-X, and BATS. We show that a high-frequency trader can profit from lead-lag relationships because of predictability, even when trading costs, latency, and execution-related risks are considered.

Keywords: Lead-lag relationship; High-frequency trading; Statistical arbitrage; Limit order book; Cross-listed stocks; Financial econometrics

1. Introduction

Lead-lag relationships have long been a subject of interest in finance, and they have been found in multiple assets and instruments.[17] But, the hypothesis that these relationships can potentially be a source of profitable statistical arbitrage is fairly recent. For example, after finding significant lead-lag relationships in NYSE stocks, Curme et al. [14] discussed the idea that lagged correlations might be exploited by a prediction model. They also believed that the resulting arbitrage opportunities may not be easily exploitable in the presence of market frictions. The same questions were also raised in Basnarkov et al. [3] in the context of foreign exchange markets. In this paper, we revisit the existence, predictability, and profitability of lead-lag relationships in detail. Our main questions are the following:

1. Can lead-lag relationships be identified in the high-frequency prices of arbitrage-linked assets?
2. If the answer to question 1 is conclusive, can returns in lagging assets be predicted?

[17] For example: stock index futures (Dimpfl and Jung [17]; Frino and West [27]), cash market and stock index futures (Chan [11]), stock and stock index futures (Brooks et al. [7]), stock index and stock index futures (Jong and Nijman [42]; Kawaller et al. [44]; Yang et al. [58]), stocks (Hou [40]), spot stock index and stock index futures markets (Herbst et al. [38]; Judge and Reancharoen [43]; Tse [57]), foreign exchange spot and futures markets (Chen and Gau [13]), and VIX markets (Bollen et al. [5])

3. If the answers to questions 1 and 2 are both affirmative, can the predictability of lagging assets be exploited by high-frequency traders (HFTers), even when important market frictions are considered?

Up to now, the profitability of statistical arbitrage from lead-lag relationships with realistic trading behavior has not been well established. Our goal is to demonstrate its economic viability by proposing a new approach based on robust lead-lag indicators, the direction probability estimation of the lagging asset's return, and the use of LOB information in an high-frequency trading (HFT) arbitrage strategy. We also consider important potential market frictions between multiple exchanges with an application to DAX 30 stocks, all of which are cross-listed in three markets: Xetra in Frankfurt, and Chi-X and BATS, both in London.

Using recent advancements in the estimation of lead-lag, stemming from Hayashi and Yoshida [37] and Hoffman et al. [39], we demonstrate that Chi-X led the high-frequency prices of most DAX 30 stocks by mere milliseconds in 2013. This surprising result is in fact in line with other studies empirically demonstrating that the most liquid, actively traded, and least expensive exchange should be the origin of price discovery. This is true in our case, since Chi-X received more quotes and trades for DAX 30 stocks on a daily basis than either Xetra or BATS. Chi-X is also the exchange with the most generous trading rebates and is thus the most competitive option for high-frequency traders, which ultimately establishes Chi-X as the price leader for the cross-listed stocks under study. We also show that all DAX 30 stocks listed at these exchanges are extremely well integrated, because their lags are limited by the speed at which information can travel. This level of precision in the estimation of cross-listed stocks lead-lag relationships has never been attained before.

Knowing that there is a definitive leader in the prices of cross-listed stocks, we then demonstrate how lagging assets returns can be predicted accurately using current and past prices observed at two exchanges. A new econometric model, the autoregressive distributed lag multinomial logistic regression, is able to utilize the existing lead-lag relationship between two price processes to predict whether the lagging asset's next return will be positive, null, or negative, with an overall accuracy exceeding 80% out-of-sample. This degree of performance is well maintained throughout our data period, further indicating the robustness of the lead-lag relationship detected in DAX 30 stocks. On our data, the proposed model's accuracy compares favorably with those of models previously suggested in the lead-lag literature, e.g., Huth and Abergel [41] and Alsayed and McGroarty [2]. It is also a significant departure from ordinary least square models, because it predicts the probabilities of the lagging asset's next return direction instead of predicting the next return itself. We show that this easier task makes it possible to build a more profitable HFT strategy by detecting more potential arbitrage opportunities with superior accuracy. Moreover, as opposed to popular frameworks based on error correction or vector autoregression models, we do not require a uniform sampling scheme of the price processes, which distinguishes our work from prior studies even further.

Fragmented markets make arbitrage opportunities more abundant for HFTers (Foucault and Biais [23]; O'Hara [48]). In this case of cross-listed stocks, whenever a lead-lag movement in a lagging asset takes longer than the information latency between exchanges to occur, an arbitrage opportunity is revealed. Earlier work on high-frequency lead-lag arbitrage failed to generate a profit due to trading costs created by market orders, with few exceptions on which we shall return. We empirically demonstrate the impossibility of profiting from the usual mid-quote signal coupled with market orders in the context of high-frequency lead-lag arbitrage. Thus, we propose a different strategy, one that makes use of limit orders, thereby reducing the exchange trading costs while also not having to pay the bid-ask spread at every arbitrage opportunity. Furthermore, the trading signal is based on level 1 prices rather than mid-quotes, leading to better-informed decisions compared to earlier studies. In a scenario where latency, trading costs, and execution-related risks are all taken into consideration, we determine that a high-frequency trader colocated at Chi-X is able to generate a net profit surpassing €1.9 million by arbitraging DAX 30 stocks in all of 2013 at only two exchanges: Xetra and BATS. The presence of market frictions dramatically impedes the trader's capacity to profit more from the detected lead-lag arbitrage opportunities, and risk management procedures are absolutely necessary to obtain a satisfying profitability.

The methodology and results in this paper are important from both the academic and practitioner standpoint. First, we contribute to the ongoing discussion about HFTers' arbitrage activities,[18] since the understanding of which is still limited in the empirical research (Chen et al. [12]). Indeed, our paper demonstrates how HFTers are realistically able to profit from a specific form of statistical arbitrage. Second, we quantify the interconnectedness of international markets in the case of cross-listed stocks by explicitly measuring the time needed between exchanges to incorporate new price information. Third, we further advance the lead-lag literature by providing the first truly profitable high-frequency lead-lag arbitrage strategy and a new econometric model that is able to predict future returns of lagging assets with an accuracy that surpasses earlier models. Furthermore, our framework is applicable to any pair of assets, making it useful for future studies on lead-lag relationships.

Our work falls under the lead-lag arbitrage literature, in which scarcely any studies have attempted to quantify the financial importance of lead-lag relationships. Alsayed and McGroarty [2]; Brooks et al. [8]; Huth and Abergel [41] are closely related to our paper, especially the first one. However, our study differs from Alsayed and McGroarty [2] on many points. Firstly, we do not work on a mid-quote basis because, as we show, this leads to suboptimal trading decisions. Each of the three papers above use that setting. We alternatively directly model the best bid and ask price processes, which allows for more precise predictions and better-informed trading decisions. Secondly, we propose an econometric model utilizing all relevant past prices observed in both the lagging and leading assets, instead of a subset of that information. Thirdly, rather than relying on liquidity-taking orders, as in the three above-mentioned papers, we employ liquidity-providing limit orders to avoid important trading costs that render all of their strategies non-viable in practice. It also allows for a more

[18] Refer to the recent Staff Report on Algorithmic Trading in U.S. Capital Markets of the SEC: https://www.sec.gov/tm/reports-and-publications/special-studies/algo_trading_report_2020 and the MiFID II Review Report on Algorithmic Trading of the ESMA: https://www.esma.europa.eu/press-news/esma-news/esma-publishes-mifid-ii-review-report-algorithmic-trading (both accessed August 12, 2022).

passive trading strategy, which we show to be profitable on our data. Finally, our application covers a new area for lead-lag arbitrage: cross-listed stocks.

The remainder of the paper is organized as follows. Section 2 introduces the literature on lead-lag relationships, where an emphasis is put on cross-listed stocks, different high-frequency arbitrage strategies, and lead-lag estimation methods in past studies. Section 3 presents the methodology used to locate and quantify lead-lag relationships. It also details the proposed econometric model in conjunction with the new HFT strategy built around it. The section ends with a description of market frictions and how we include them into our estimations. Section 4 is dedicated to the data from Xetra, Chi-X, and BATS, and also presents the latencies and costs we utilize. Section 5 analyzes the empirical results of our methodology and discusses their implications. Section 6 concludes the paper.

2. Literature review

As discussed in the introduction, lead-lag relationhips have been observed in most financial assets and instruments. The particular case of cross-listed stocks has been studied at an intraday frequency in Grammig et al. [31]; Pascual et al. [49]; Frijns et al. [24]; Frijns et al. [25]; Ghadhab and Hellara [28]; and Frijns et al. [26]. They all analyze cross-listed stock price discovery based on variations of Hasbrouck's information shares (Hasbrouck [33]) and/or the component shares of Gonzalo and Granger [30]. Grammig et al. [31] sample 10-second intervals of mid-quote prices of three German firms cross-listed in New York (NYSE) and Frankfurt (Xetra) from August to October 1999, and find that price discovery mostly originated from the home exchange. Pascual et al. [49] arrive at the same conclusion in the case of five Spanish ADRs listed on the NYSE and SSE at a one-minute resolution in 2000, as do Frijns et al. [24] on four Australian and five New Zealand firms from 2002 to 2007 at a minute level. Ghadhab and Hellara [28] also corroborate the idea that local markets are dominant for cross-listed stocks, but find that foreign markets contribute more to price discovery for multiple-listed firms, even more so when their trading costs are lower. Other factors affect the origin of price discovery for cross-listed stocks. Indeed, Frijns et al. [25]

suggest that a reduced bid-ask spread and a higher trade activity, small trades in particular, have a positive and causal impact on price discovery, from a sample of cross-listed Canadian stocks in the US from 1996 to 2011, at a minute frequency. These recur in Frijns et al. [26], which finds a bilateral causality between liquidity in an exchange and its contribution to price discovery. These authors also obtain that algorithmic activity is negatively related to price discovery for Canadian cross-listed stocks in the US from 2004 to 2017. None of the papers mention the possibility of an arbitrageur exploiting these lead-lag relationships, nor do they measure how predictable the lagging assets returns are. We aim to answer these questions by proposing a novel HFT strategy and a new econometric model for cross-listed stocks. Our methodology also considers important limiting factors of arbitrage, mainly, trading costs, latency, and execution-related risks. The proposed model is also computationally simple enough to be used by HFTers in practice.

Very few papers have tried to develop arbitrage strategies or predictive models based on the concept of lead-lag, and none in the context of cross-listed stocks: Judge and Reancharoen [43] and Li et al. [32] use daily data; Brooks et al. [8] and Stübinger [56] focus on uniformly sampled intraday data; and Huth and Abergel [41] and Alsayed and McGroarty [2], the closest studies to our paper, also use LOB data. Brooks et al. [8] investigate the lead-lag relationship between the spot index and futures contract of the FTSE 100 at a 10-minute frequency. They are able to predict, one step ahead, the direction of the return in the lagging spot price, with an out-of-sample accuracy approaching 70%, based on a version of the error correction model (ECM) of Engle and Granger [18]. Nonetheless, because of trading costs, their round-trip trade strategy is unable to outperform a passive buy-and-hold strategy. In the same vein, Huth and Abergel [41] are also not able to profit from the lead-lag relationship they detect in a futures-stock pair, since paying the bid-ask spread at every opportunity is too expensive. Even though their linear regression model predict the next mid-quote return at the next trade of the lagging stock with an accuracy of 60%, the opportunities detected do not cover the market orders costs.

On the other hand, Stübinger [56] and Alsayed and McGroarty [2] find economically significant profit-generating strategies by exploiting lead-lag relationships. Stübinger [56] proposes the "optimal causal path algorithm" to uncover the lead-lag structure between two time series, and then applies it to S&P 500 constituents at a minute level, to identify promising stocks for a pair trading–type strategy. The strategy limits excessive trading by only selecting statistically high returns of the leading stock that also cover the trading costs of market orders. Positions are closed after ℓ minutes, where ℓ is the lag estimated from the optimal causal path algorithm. This trading signal allows the author to significantly outperform a buy-and-hold strategy of the S&P 500 index after transaction costs. But, in a high-frequency setting where lag is measured in milliseconds, as in our study, the trading signal of Stübinger [56] would result in an insignificant number of trades, since price movements at that scale seldom cover the bid-ask spread. Alsayed and McGroarty [2] profit from lead-lag arbitrage across international futures with a new forecasting framework yielding over 85% accuracy in lagging contracts' mid-quote changes. Their framework is based on the concept of clusters, which are uninterrupted, contiguous observations of prices that allow them to predict mid-quote movements and trade at a high frequency. But, we question the strategy's practical profitability because their profit calculations use mid-quote returns and not actual execution prices. We are proposing a novel high-frequency strategy relying on limit orders to circumvent the profitability issues of earlier studies. Our practical methodology also gets as close as possible to real-life HFT, thus making our results more concrete and accurate. In both Huth and Abergel [41] and Alsayed and McGroarty [2], the leading asset leads by mere fractions of a second: around 300 milliseconds in the former and down to 25 milliseconds for a particular pair in the latter. This highlights the importance of newer methodologies enabling sub-second lead-lag estimation.

Considering that today's integrated markets rely heavily on advanced information technology to connect traders and exchanges around the globe, aggregated data at the minute level is not suitable to uncover lead-lag relationships between cross-listed stocks. This is especially true when exchanges are geographically close. As shown in Budish et al. [10], the correlation of related instruments only breaks down at a millisecond resolution in well-integrated markets, even though their correlation seem nearly perfect at a minute level. But, using sub-second data, i.e., trades and quotes (TAQ) from LOB data, to quantify lead-lag relationships has its challenges: it is neither synchronously nor regularly observed. As noted in Hayashi and Yoshida [37] and Zhang [59], among others, earlier estimators based on previous-tick interpolation are severely biased whenever the processes are not synchronously observed. This is true for Granger's causality (Granger [32]) and for Hasbrouck's information share (Hasbrouck [33]) models when working with HFT data, because correlation estimates decrease when the processes are synchronously sampled at high frequencies. This downward correlation bias effect was first studied in Epps [19]. Furthermore, if the two processes differ in noise, microstructure frictions, or liquidity, these methods will not be consistent (Putniņš [53]). Hasbrouck [34] extends the vector error correction model (VECM) to accomodate ultra high-frequency data, resulting in better information share estimations and opening the door to causal methods on LOB data (see the first paragraph of this section for the overview of lead-lag papers based on causality). The author finds significant differences in the information share estimation from uniformly sampled data and LOB event data, a result that needs further investigation. Since 2010, some consistent estimators of lead-lag at a high frequency have been proposed based on LOB event time and correlation methods (e.g., Hayashi and Koike [36]; Hoffman et al. [39]), and on large trade events (Pomponio and Abergel [50]), making it possible to depart from previous-tick interpolation and uniform sampling, to instead use the LOB as is. We are the first to investigate lead-lag relationships of cross-listed stocks at that level of precision. Being able to work at the sub-second horizon is absolutely necessary in our case, because the geographical proximity of the exchanges allows information to flow between them nearly instantly.

3. Methodology and framework

We introduce the ideas behind the results presented in Section 5. Even though our application covers cross-listed stocks, the general methodology and framework in this section are applicable to any financial market where a high-frequency trader suspects that a lead-lag relationship exists between any pair of assets.

Subsection 3.1 details how we find lead-lag relationships between processes and how to quantify their strength. Subsection 3.2 proposes an econometric model able to exploit an existing lead-lag relationship by predicting the lagging process' future directional movements from past information on the leading process. Subsection 3.3 presents an HFT strategy created from the econometric model predictions. Finally, subsection 3.4 is dedicated to the market frictions we consider when computing our trading profits.

3.1. Lead-lag relationships

There are two main schools of thought as regards the ways of mathematically defining and detecting lead-lag relationships, or the price discovery origin: causality methods (e.g., Granger [32]) and correlation methods (e.g., Herbst et al. [38]). The latter approach makes it possible to explicitly measure the timing relationship between time series, which provides valuable information in a trading context. Following that literature, there exists a lead-lag relationship in a pair of stochastic processes ($\{X_t\}$,$\{Y_t\}$) with observations ($\{x_t\}, \{y_t\}$) whenever their cross-correlation with lag ℓ, $\operatorname{Corr}(X_t, Y_{t+\ell})$, is statistically different from 0 for any $\ell \neq 0$. The optimal lag ℓ^* is defined as

$$\ell^* \equiv \underset{\ell \in \mathbb{R}}{\arg\max} \, |\operatorname{Corr}(X_t, Y_{t+\ell})| = \underset{\ell \in \mathbb{R}}{\arg\max} \, |\rho_{X,Y}(\ell)|,$$

where $\rho_{X,Y}(\ell)$ is the lagged Pearson correlation coefficient $\rho_{X,Y}(\ell) \equiv \dfrac{\operatorname{Cov}(X_t, Y_{t+\ell})}{\sqrt{\operatorname{Var}(X_t)\operatorname{Var}(Y_{t+\ell})}}$, $\operatorname{Cov}(X_t, Y_{t+\ell})$ is the lagged cross-covariance of processes ($\{X_t\}$,$\{Y_t\}$), and $\operatorname{Var}(\cdot)$ is their variance. Whenever $\ell^* \neq 0$, the relationship between $\{X_t\}$ and $\{Y_t\}$ is not contemporaneous and it establishes that there is lead-lag between the processes. When $\ell^* > 0$, $\{X_t\}$ leads $\{Y_t\}$ and vice versa for $\ell^* < 0$. Knowledge of the leader at t can potentially be exploited to forecast the lagging process at $t + \ell^*$.

In this paper, we rely on high-frequency data, which is notable for being non-synchronous and irregularly observed. "Non-synchronous" means that the two processes are observed at different times, and "irregularly observed" refers to irregular intervals between observation times of the processes. These features drive us to depart from older lead-lag estimation methods used in the literature, as mentioned earlier in Section 2. Hayashi and Yoshida [37] propose a covariance estimator for non-synchronous and irregularly observed diffusion processes, resulting in the following consistent cross-correlation estimator:

$$\hat{\rho}^{HY}_{X,Y} = \frac{\sum_i \sum_j \Delta X(I_i^X)\Delta Y(I_j^Y)\mathbb{I}_{\{I_i^X \cap I_j^Y \neq \emptyset\}}}{\sqrt{\sum_i [\Delta X(I_i^X)]^2 \sum_j [\Delta Y(I_j^Y)]^2}},$$

where

$$\mathbb{I}_{\{A\}} = \begin{cases} 1, & \text{if A is true,} \\ 0, & \text{if A is false} \end{cases}$$

is the indicator function. The processes ($\{X_t\}$,$\{Y_t\}$) have discrete observation times $0 = t_1^X < t_2^X < \cdots < t_n^X = T^X$ and $0 = t_1^Y < t_2^Y < \cdots < t_m^Y = T^Y$ with intervals $I_i^X = (t_{i-1}^X, t_i^X]$, $I_j^Y = (t_{j-1}^Y, t_j^Y]$ and $\Delta X(I_i^X) = x_{t_i^X} - x_{t_{i-1}^X}$, $\Delta Y(I_j^Y) = y_{t_j^Y} - y_{t_{j-1}^Y}$. Hoffman et al. [39] extend this estimator to include the lag ℓ:

$$\hat{\rho}^{HY}_{X,Y}(\ell) = \frac{\sum_i \sum_j \Delta X(I_i^X)\Delta Y(I_j^Y)\mathbb{I}_{\{I_i^X \cap (I_j^Y)_\ell \neq \emptyset\}}}{\sqrt{\sum_i [\Delta X(I_i^X)]^2 \sum_j [\Delta Y(I_i^Y)]^2}}$$

where $(I_j^Y)_\ell = (t_{j-1}^Y + \ell, t_j^Y + \ell]$. This makes it possible to obtain a practical and unbiased estimation of ℓ^* on HFT data:

$$\widehat{\ell^*} = \underset{\ell \in \mathbb{R}}{\arg\max}\, |\hat{\rho}^{HY}_{X,Y}(\ell)|,$$

which is the estimator used in this paper. In order to quantify the overall side and strength of the lead-lag relationship, Huth and Abergel [41] introduce the Lead-Lag Ratio (LLR) measuring the asymmetry of the cross-correlation function:

$$LLR_{X,Y} \equiv \frac{\sum_{g \in \mathcal{G}} \hat{\rho}^{HY}_{X,Y}(\ell_g)^2}{\sum_{g \in \mathcal{G}} \hat{\rho}^{HY}_{X,Y}(-\ell_g)^2}$$

for $\{\ell_g \mid g \in \mathcal{G}\}$, a discrete time grid of positive lags. Whenever $LLR_{X,Y} > 1$, $\{X_t\}$ leads $\{Y_t\}$ and the higher $LLR_{X,Y}$ is, the more $\{X_t\}$ leads $\{Y_t\}$. This statistic is also applied to detect lead-lag relationships in our data.

3.2. Econometric model

We concentrate on the models of Huth and Abergel [41] and Alsayed and McGroarty [2] since they are the only studies whose methodologies are directly developed on unsampled LOB data. Huth and Abergel [41] are predicting the direction of the mid-quote move (up or down) at the next trade of the lagging mid-quote process $\{Y_t\}$ by taking the sign of a linear combination that uses the leader's past mid-quote moves as the only exogenous variables, like so:

$$\widehat{R}^Y_j \equiv \text{sign}(\widehat{\Delta Y}(I^Y_j)) = \text{sign}\left(\sum_{k=1}^{p} \beta_k \sum_{i:t^X_i<t^Y_{j-1}} \Delta X(I^X_i)\mathbb{I}_{\{I^X_i \cap (I^Y_j)_{\ell_k} \neq \emptyset\}}\right)$$

where p is the last statistically significant lag. They set $\beta_k = \hat{\rho}^{HY}_{X,Y}(\ell_k)$ and achieve around 60% directional accuracy on test days. The model's core idea is a binary classification, when in fact, a logistic regression would be more appropriate than taking the sign of a model that is designed for a harder prediction problem. Predictions that fall close to 0 can also be problematic since they lie around the model's decision boundary, where predictions are most uncertain (Nguyen et al. [47]). Adding a null prediction seems necessary for HFT whenever that occurs. Null predictions have been considered in the next contribution.

Alsayed and McGroarty [2] define *clusters* as sets of contiguous process variations uninterrupted by variations of a second process observed in parallel. They define $\left\{C^X_{i,n} \;\middle|\; i,n \in \mathbb{N}^+\right\}$ as the set of clusters of process $\{X_t\}$, where the subscript i refers to the cluster index and n the variation index within each cluster. The same definition holds for process $\{Y_t\}$. Figure 14 illustrates the concept of clusters.

Suppose that $\{X_t\}$ leads $\{Y_t\}$, and define $\overline{C}^{\cdot}_{i,n}$ as the mid-quote returns of both processes, Alsayed and McGroarty [2] predict the next cluster's direction of the lagging asset, $R_{\overline{C}^Y_i} \equiv$

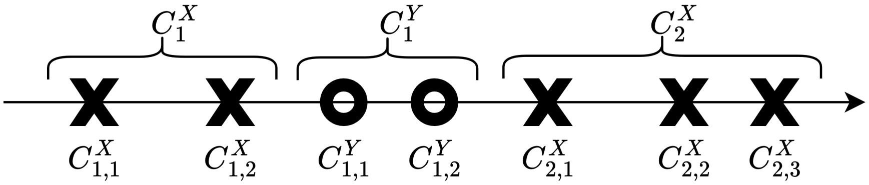

Fig. 14. Time-line illustration of dual process clusters. Observations of process $\{X_t\}$ are marked by an "X" and those of $\{Y_t\}$ are marked by an "O." Taken from Alsayed and McGroarty [2].

$\text{sign}\left(\sum_n \overline{C}_{i,n}^Y\right)$, with the following rule:

$$\widehat{R}_{\overline{C}_i^Y} = \begin{cases} +1, & \text{if} \max_n \left(\overline{C}_{i,n}^X\right) \geq K^{AM} \\ -1, & \text{if} \min_n \left(\overline{C}_{i,n}^X\right) \leq -K^{AM} \\ 0, & \text{otherwise,} \end{cases}$$

where $K^{AM} \in \mathbb{R}_0^+$ is a preset threshold. They achieve a directional accuracy in excess of 85% on pairings of S&P 500, FTSE 100, and DAX futures contracts in 2012. This high level of accuracy can be explained by the high $LLR_{X,Y}$ in the three asset pairs studied. Only relying on the leader's latest cluster might be hazardous for asset pairs with a weaker lead-lag relationship.

Huth and Abergel [41] and Alsayed and McGroarty [2] both offer interesting predictive models that are able to exploit HFT lead-lag relationships in their respective financial contexts. The use in Huth and Abergel [41] of the leading process' past relevant information, the simplicity of the Alsayed and McGroarty [2] model, and the trader's ability to set a confidence threshold are all important qualities in HFT econometric models. We extend their contributions by proposing a model that takes into account the aforementioned overlooked aspects. Following Alsayed and McGroarty [2], we set clusters of the leading price process as $C_i^X = \left\{C_{i,j}^X \;\middle|\; j = 1,\ldots,n_i^X \in \mathbb{N}^+\right\}$ and the lagging price process' as $C_i^Y = \left\{C_{i,j}^Y \;\middle|\; j = 1,\ldots,n_i^Y \in \mathbb{N}^+\right\}$, where $i = 1,2,\ldots,N$ for N the number of clusters, and $C_{i,j}^X = \Delta X\left(I^X_{\sum_{k<i} n_k^X + j}\right)$, $C_{i,j}^Y = \Delta Y\left(I^Y_{\sum_{k<i} n_k^Y + j}\right)$ the absolute variations of the two

price processes (any price process, not necessarily mid-quote). We define $r_{C_i^X} = \sum_{j=1}^{n_i^X} C_{i,j}^X$ as the total price process variation within cluster C_i^X and the same definition applies for $\{Y_t\}$. Without loss of generality, we assume that the first cluster we observe is from $\{X_t\}$, and the last one is from $\{Y_t\}$. We are interested in predicting the direction of $r_{C_i^Y}$ based on past observations of ($\{X_t\}$,$\{Y_t\}$), i.e.,

$$R_{C_i^Y} \equiv \text{sign}\left(r_{C_i^Y}\right) = \begin{cases} +1, & \text{if } r_{C_i^Y} > 0 \\ 0, & \text{if } r_{C_i^Y} = 0 \\ -1, & \text{if } r_{C_i^Y} < 0. \end{cases}$$

To do so, we propose the autoregressive distributed lag multinomial logistic regression (ADLMLR) to model $R_{C_i^Y}$. It generalizes the logistic models for autoregressive binary variables introduced in Bonney [6] in two ways. Firstly, it departs from a binary dependent variable to a multicategorical one, allowing for the modeling of a larger spectrum of systems. Secondly, $\{Y_t\}$ is not only autoregressive, it is autoregressive with a distributed lag for $\{X_t\}$, thus incorporating past values of both processes. Our model is an important departure from conventional approaches based on error correction models (ECM) (for example, Brooks et al. [8]; Engle and Granger [18]; Frijns et al. [24]; Hasbrouck [33]; Judge and Reancharoen [43]; Pascual et al. [49]; Yang et al. [58]) or vector autoregressive models (VAR) (see Dimpfl and Jung [17]; Hou [40]) since it does not require the processes to be synchronously and regularly observed in time, thanks to the use of clusters. We also depart from an ordinary least squares (OLS) framework to a probabilistic one, where we are interested in predicting the probability of the class of the next return's direction (positive, neutral, or negative) instead of quantifying the return itself. This probabilistic task is easier to accomplish, hence the model predictions are more robust. As we will show, this leads to a greater profitability potential when incorporated into an HFT strategy.

The proposed ADLMLR model for $R_{C_i^Y}$ is as follows. Supposing a (auto)dependence lag of order $D \in \mathbb{N}^+$,

$$\Big(R_{C_i^Y} \mid r_{C^Y_{i-D:i-1}}, r_{C^X_{i-D+1:i}}\Big) \sim \text{Multinoulli}(p_{i,-1},\, p_{i,0},\, p_{i,+1}),$$

where $p_{i,\cdot} \in [0{,}1]$, $\sum p_{i,\cdot} = 1\ \forall i$, are the conditional probabilities of their respective return direction based on the past observations of $(\{X_t\},\{Y_t\})$ and are denoted by:

$$\begin{aligned} p_{i,-1} &= P\Big(R_{C_i^Y} = -1 \mid r_{C^Y_{i-D:i-1}}, r_{C^X_{i-D+1:i}}\Big), \\ p_{i,0} &= P\Big(R_{C_i^Y} = 0 \mid r_{C^Y_{i-D:i-1}}, r_{C^X_{i-D+1:i}}\Big), \\ p_{i,+1} &= P\Big(R_{C_i^Y} = +1 \mid r_{C^Y_{i-D:i-1}}, r_{C^X_{i-D+1:i}}\Big), \end{aligned}$$

for $r_{C_{i-D:i}} = \{r_{C_{i-D}}, r_{C_{i-D+1}}, \ldots, r_{C_{i-1}}, r_{C_i}\}$. We define the conditional probabilities from the *logit* function with an autoregressive distributed lag-like model:

$$\begin{aligned} \ln\left(\frac{p_{i,-1}}{p_{i,+1}}\right) &= \alpha_{-1} + \sum_{j=0}^{D-1} \beta_{j,-1} r_{C^X_{i-j}} + \sum_{j=1}^{D} \gamma_{j,-1} r_{C^Y_{i-j}}, \\ \ln\left(\frac{p_{i,0}}{p_{i,+1}}\right) &= \alpha_0 + \sum_{j=0}^{D-1} \beta_{j,0} r_{C^X_{i-j}} + \sum_{j=1}^{D} \gamma_{j,0} r_{C^Y_{i-j}}. \end{aligned}$$

Since we also have $\sum p_{i,\cdot} = 1$, we can find the conditional probabilities:

$$\begin{aligned} p_{i,-1} &= \frac{e^{\theta_{i,-1}}}{1 + e^{\theta_{i,-1}} + e^{\theta_{i,0}}}, \\ p_{i,0} &= \frac{e^{\theta_{i,0}}}{1 + e^{\theta_{i,-1}} + e^{\theta_{i,0}}}, \\ p_{i,+1} &= \frac{1}{1 + e^{\theta_{i,-1}} + e^{\theta_{i,0}}}, \end{aligned}$$

where

$$\begin{aligned} \theta_{i,-1} &= \alpha_{-1} + \sum_{j=0}^{D-1} \beta_{j,-1} r_{C^X_{i-j}} + \sum_{j=1}^{D} \gamma_{j,-1} r_{C^Y_{i-j}}, \\ \theta_{i,0} &= \alpha_0 + \sum_{j=0}^{D-1} \beta_{j,0} r_{C^X_{i-j}} + \sum_{j=1}^{D} \gamma_{j,0} r_{C^Y_{i-j}}. \end{aligned}$$

The parameters of the model $\Theta = \{\alpha_{-1}, \alpha_0, \beta_{0,-1}, \dots, \beta_{D-1,-1}, \beta_{0,0}, \dots, \beta_{D-1,0}, \gamma_{1,-1}, \dots, \gamma_{D,-1}, \gamma_{1,0}, \dots, \gamma_{D,0}\}$ are found by maximum likelihood estimation of

$$\mathcal{L}(\Theta) = \prod_{i=D}^{N} (p_{i,-1})^{\mathbb{I}\left\{R_{C_i^Y}=-1\right\}} (p_{i,0})^{\mathbb{I}\left\{R_{C_i^Y}=0\right\}} (p_{i,+1})^{\mathbb{I}\left\{R_{C_i^Y}=+1\right\}},$$

for N the number of clusters in $\{Y_t\}$, so that

$$\widehat{\Theta} = \underset{\Theta \in \mathbb{R}^{4D+2}}{\arg\max}\, \mathcal{L}(\Theta).$$

We use the BFGS algorithm of Broyden [9]; Fletcher [22]; Goldfarb [29]; Shanno [55] to solve for $\widehat{\Theta}$. The largest predicted probability in vector $\widehat{\mathbf{p}}_i = [\widehat{p}_{i,-1}\ \widehat{p}_{i,0}\ \widehat{p}_{i,+1}]$ determines the direction of the total variation in cluster C_i^Y:

$$\widehat{R}_{C_i^Y} = \begin{cases} +1, & \text{if } \max(\widehat{\mathbf{p}}_i) = \widehat{p}_{i,+1},\ \widehat{p}_{i,+1} \geq K \\ -1, & \text{if } \max(\widehat{\mathbf{p}}_i) = \widehat{p}_{i,-1},\ \widehat{p}_{i,-1} \geq K \\ 0, & \text{otherwise,} \end{cases}$$

where $K \in [0,\ 1]$ is a preset decision threshold controlling the minimum confidence needed to make a prediction. Its empirical selection will be discussed in Section 5.2. The ADLMLR model is also closely related to the autoregressive conditional multinomial-autoregressive conditional duration (ACM-ACD) model of Russell and Engle [54] since price changes are also assumed to follow a multinomial logistic model. But in their case, only a single asset is considered. ADLMLR alters the ACM part of their model to directly consider related assets by also conditioning on the price changes of a leading asset, thus adding a distributed lag process in the autoregression of the lagging asset. Our price direction function is also more adapted to trading than that of Russell and Engle [54] because of the added confidence threshold K.[19]

[19] Note that we do not consider the duration effect studied in Russell and Engle [54], which can be further investigated in a subsequent paper. We choose to focus on the price dynamics for now.

3.3. High-frequency arbitrage strategy

With market orders, Brooks et al. [8] and Huth and Abergel [41] are not able to profit from their predictions, as paying the bid-ask spread at every opportunity is prohibitive for a HFTer, even more so considering exchange trading costs. Predicting the direction of mid-quote movement is also not the most practical way of building an HFT strategy since orders cannot be executed at that price — another problem discussed in Huth and Abergel [41]. To circumvent these issues, we are predicting the return direction in the best bid and best ask prices based on the econometric model introduced in the previous subsection. In other words, a first model instance is used for the best bid price process and a second one is dedicated to the best ask. We are also relying on limit orders to reduce trading costs.

We assume an existing lead-lag relationship between a leader $\{X_t^{Bid/Ask}\}$ and a lagging process $\{Y_t^{Bid/Ask}\}$, which are the best bid/ask price processes. We also assume that our econometric model is able to utilize that relationship to generate adequate predictions. Based on these assumptions, we are interested in profiting from the predicted directions in clusters of $\{Y_t^{Bid/Ask}\}$: $\widehat{R}_{C_i^{Y^{Bid/Ask}}}$. For a tick size of δ, the novel HFT strategy is as follows:

- Bid price process:
  - When $\widehat{R}_{C_i^{Y^{Bid}}} = -1$, do all actions at the same time:
    1. Send a marketable sell limit order of volume V_i^{Bid} at the current value of $\{Y_t^{Bid}\}$;
    2. Send a buy limit order of volume V_i^{Bid} at the current value of $\{Y_t^{Bid}\}$ minus δ;
    3. Send a stop buy limit order of volume V_i^{Bid} with stop and limit prices equal to the current value of $\{Y_t^{Bid}\}$ plus 2δ.
  - When $\widehat{R}_{C_i^{Y^{Bid}}} \in \{0, 1\}$: do nothing.
  - When a position has been open for M minutes, send a market buy order to close.
- Ask price process:
  - When $\widehat{R}_{C_i^{Y^{Ask}}} = 1$, do all actions at the same time:

1. Send a marketable buy limit order of volume V_i^{Ask} at the current value of $\{Y_t^{Ask}\}$;
2. Send a sell limit order of volume V_i^{Ask} at the current value of $\{Y_t^{Ask}\}$ plus δ;
3. Send a stop sell limit order of volume V_i^{Ask} with stop and limit prices equal to the current value of $\{Y_t^{Ask}\}$ minus 2δ.

– When $\widehat{R}_{C_i^{Y^{Ask}}} \in \{-1,\ 0\}$: do nothing.

– When a position has been open for M minutes, send a market sell order to close.

A short (long) position is open when the marketable sell (buy) limit order hits the market and the buy (sell) limit order tries to close it whenever the lagging process $\{Y_t^{Bid}\}$ ($\{Y_t^{Ask}\}$) moves in the predicted direction. This allows us to capture a potential profit of δ when our econometric model makes a good prediction. No new position is open until the previous one has been closed. The stop limit orders are employed for risk management in the case of a wrong prediction; the same goes for closing market orders. Additional details of the strategy are presented in Appendix A.

3.4. Market frictions: latency, risks, and costs

In order to be as practical as possible, we use the Deltix QuantOffice trading software suite. This software only manages back-office operations and replays the LOB messages for backtesting purposes, letting us to get closer to real-life high-frequency trading. When using historical data, like in this paper, QuantOffice emulates the trades and quotes in a live-streaming environment, and computes the profit and loss results at the end of a strategy's execution. It is possible to bypass the software and implement an equivalent testing program, but we utilize the professional suite to ensure the quality of the results.[20]

[20] Deltix has worked in collaboration with more than 100 banks, brokers, institutional investors and high-frequency traders in the U.S. Hence, the results provided in this paper are representative of what would have been obtained in these institutions before switching the strategy to live markets directly from the same code implemented in QuantOffice. They are the closest results to real time trading possible in an academic context.

Latency is of paramount importance in HFT, as shown in Poutré et al. [52]. So, we use a simplified version of their methodology to account for latency in our empirical results. By latency, we mean the total time it takes for a trader to interact with the market when new information arrives. Hasbrouck and Saar [35] measure latency on three components: the time it takes for a trader to learn about an event, to generate a response, and for the exchange to act on that response. Considering a HFTer colocated at the leading exchange, the first two components of latency are the amount of time required for information generated at a lagging exchange to arrive and its treatment by the HFTer's server and trading algorithm. This is due to the finite speed of light causing a delay in the observed LOB between the source of information (lagging exchange) and its point of observation (leading exchange). To replicate that relativistic effect for a HFTer, we wait for an amount of time equal to the true one-way information transportation time plus its treatment time before entering the lagging exchange's data into the HFT strategy, thus delaying it. So, for a HFTer colocated at the leading exchange, it is as if its trading algorithm only observes past LOB states of the geographically distant lagging exchange, as it would in practice. Moving forward, this will be referred to as the first half of latency.

The last component of latency, which we will refer to as the second half of latency, is treated similarly. When the HFT strategy of Section 3.3 generates a trade signal, the orders are only sent to the execution engine after a time delay that corresponds to the same one-way information transportation time between exchanges, plus the receiving exchange's matching engine delay. So, a HFTer cannot interact infinitely rapidly with a geographically distant lagging exchange, as is the case in practice. For convenience, we assume that the HFT server is able to process a stream of level 1 data with the same efficiency as an exchange server. This allows us to use the same total latency value for the first and second halves of latency. In the next section, Table 20 presents the latency values employed.

The high-frequency strategy is exposed to both execution and non-execution risks since it utilizes market and limit orders. Those risks are taken into account using a set of fixed professional rules determining if, when, and at what price the orders sent would have been executed in practice, making sure to always respect the price-time priority rules. The complete list of execution rules is given in Appendix C.[21] We also compute exchange trading costs after an order's execution, which are shown in Table 20 of the next section. Liquidity removal costs for marketable limit and market orders, and liquidity-providing costs for limit orders are taken into account.

4. Data

DAX 30 (which was extended to DAX 40 on November 24, 2020) is a German stock index containing 30 of the country's largest blue chip companies. Table 18 lists its constituents in 2013, and Table 19 details some of their stylized facts. Xetra, operated by Deutsche Börse AG at the Frankfurt Stock Exchange, is the reference order-driven trading venue for German stocks and has normal trading hours of 9:00 a.m. to 5:30 p.m. CET.[22] Chi-X Europe, also an order-driven exchange, is a cost-effective pan-European alternative to the largest European exchanges, with continuous trading hours between 8:00 a.m. and 4:30 p.m. GMT, located in London. Finally, BATS Europe (Better Alternative Trading System) is another London–based pan-European stock exchange, founded in 2008. BATS Europe was a direct competitor to Chi-X Europe, with the same normal trading hours, but it ultimately acquired the latter in 2011. The London–based exchanges lag one hour behind Xetra because of different time zones, but all their normal trading hours overlap completely, from opening to closing.

Our data covers DAX 30 stocks in the three European exchanges listed above: Xetra, Chi-X, and BATS, and spans six months in 2013, from February to July, inclusively, thus covering 125 trading days. The data of Chi-X and BATS was acquired from BEDOFIH (*Base Européenne de Données Financières à Haute Fréquence*) and it contains the trades and quotes at a millisecond precision for the first 20 LOB levels but only the first level is

[21] The execution rules were obtained from Deltix.

[22] Xetra offers the "continuous trading with auctions" service for its more liquid securities. Call auctions occur three times in a regular trading day for DAX 30 stocks: from 8.50 am to 9.00 am at the earliest (opening auction), from 1:00 p.m. to 1:02 p.m. at the earliest (intraday auction), and from 5:30 p.m. to 5:35 p.m. at the earliest (closing auction), with random end times. Continuous trading occurs in between auctions and only these periods are used in our study. See https://www.xetra.com/xetra-en/trading/trading-models/continuous-trading-with-auctions for the detailed trading models of Xetra.

used in this study. Microsecond precision data for these two exchanges only started in 2019.[23] Xetra's raw data contains every market event sent by the exchange, and we use the Xetra Parser software of Bilodeau [4] to rebuild the first level of the LOB at a microsecond precision for each update. The timestamps are then rounded to the nearest greater millisecond, for use in conjunction with the previous data sets. Rounding Xetra's timestamps is absolutely necessary to keep the most intact sequential order of events between the three exchanges, given that both Chi-X's and BATS's timestamps have a lower precision level. The light travel time between Frankfurt and London, i.e., the absolute physical limit in information latency, is north of 2 ms, so analyzing the lead-lag relationships of Xetra's stocks with smaller, more precise latency values is unnecessary. Moreover, because the average rate of events is around one per 135 ms for the most active stock (Chi-X:DBK), very few events occur in the same millisecond at multiple exchanges. Hence, overall, losing precision on Xetra's data by rounding its timestamps to a millisecond accuracy does not significantly alter the lead-lag relationship analysis, nor the strategy's results reported in this study.

Panel A of Table 20 details the latencies used to generate the strategies' results, which have been communicated directly by the exchanges and a trading connectivity firm.[24] Total latencies are rounded to the nearest non-zero integer. Note that microwaves began to be adopted in 2010 by HFTers and that a one-way trip between London and Frankfurt was around 2.3 ms with that technology. Thus, the total latency used for both links in this study are convervative, meaning that the profitability results presented in Section 5.3 are also conservative, since our strategy most probably reacts more slowly than the most sophisticated HFTers at that time. Panel B shows the trading costs of the three exchanges in 2013,[25] and

[23] See the European Financial data Institute's data description guide for the complete documentation (accessed April 6, 2023).

[24] Note that Chi-X and BATS servers are located in Equinix Slough (LD4), 32 km west of Central London, and Xetra servers are in Frankfurt (FR2). Also note that one-way transportation latency is half of a round trip. Sources used are: https://www.marketsmedia.com/extent-of-adoption-of-microwave-technology-in-europe-revealed (Chi-X/Xetra one-way on fiber optics to be conservative), Deutsche Börse Group [15] (Xetra exchange latency), and https://cdn.cboe.com/resources/press_releases/BATS_Europe_Latency_Update_FINAL.pdf (BATS exchange latency).

[25] Deutsche Börse Group [16] contains the trading costs of DAX stocks at Xetra, and https://www.cboe.com/europe/equities/notices/41029/fee_schedule/ the trading costs of Chi-X and BATS. All trading costs are effective January 2, 2013.

Table 18. DAX 30 constituents from February to July 2013.

Ticker	Company	Prime Standard Sector
ADS	Adidas	Consumer
ALV	Allianz	Insurance
BAS	BASF	Chemicals
BAYN	Bayer	Chemicals
BEI	Beiersdorf	Consumer
BMW	BMW	Automobile
CBK	Commerzbank	Banks
CON	Continental	Automotive
DAI	Daimler AG	Automobile
DB1	Deutsche Börse	Financial Services
DBK	Deutsche Bank	Banks
DPW	Deutsche Post	Transportation & Logistics
DTE	Deutsche Telekom	Telecommunication
EOAN	E.ON	Utilities
FME	Fresenius Medical Care	Pharma & Healthcare
FRE	Fresenius	Pharma & Healthcare
HEI	HeidelbergCement	Construction
HEN3	Henkel	Consumer
IFX	Infineon Technologies	Technology
LHA	Deutsche Lufthansa	Transportation & Logistics
LIN	Linde	Chemicals
LXS	Lanxess	Chemicals
MRK	Merck	Pharma & Healthcare
MUV2	Munich Re	Insurance
RWE	RWE	Utilities
SAP	SAP	Software
SDF	K+S	Chemicals
SIE	Siemens	Industrial
TKA	Thyssenkrupp	Industrial
VOW3	Volkswagen AG	Automobile

Panel C documents the rules used by Xetra to determine stocks' tick sizes.[26] Chi-X and BATS subsequently use the same tick sizes for cross-listed stocks also traded at Xetra.

[26] https://www.xetra.com/xetra-en/trading/trading-models/trading-parameter-tick-size. All websites referenced in this section were accessed on September 7, 2022.

Table 19. Stylized facts of the DAX 30 stocks from February to July 2013.

		Xetra			Chi-X			BATS		
Ticker	Market Cap ($B)	Daily Trades	Daily Quotes	Daily Volume	Daily Trades	Daily Quotes	Daily Volume	Daily Trades	Daily Quotes	Daily Volume
ADS	18.62	4 065.37	45 892.78	717 000.01	4 754.04	90 547.10	351 458.26	926.17	29 146.06	58 485.46
ALV	62.80	4 750.37	63 093.81	1 568 537.26	4 812.42	83 739.54	498 074.92	2 079.18	45 718.41	166 271.06
BAS	86.42	7 924.10	95 070.57	2 481 711.35	9 282.54	170 845.96	1 038 437.31	2 434.26	77 506.34	196 013.78
BAYN	78.62	6 687.76	76 045.05	1 661 952.61	10 481.14	157 124.54	935 971.42	1 764.42	58 777.43	127 529.10
BEI	15.28	2 196.01	35 603.93	379 020.22	2 721.87	63 721.97	202 624.60	432.15	22 821.12	25 154.20
BMW	50.68	5 919.62	73 191.91	1 483 157.74	7 250.10	122 096.66	651 476.35	1 484.31	52 993.49	133 672.82
CBK	9.10	7 638.83	64 760.41	30 018 813.68	4 709.13	93 422.61	4 263 311.16	1 221.26	37 149.70	951 537.95
CON	19.05	3 224.14	44 399.19	429 689.75	3 124.57	76 207.83	167 750.59	806.37	25 400.49	42 794.27
DAI	48.04	9 351.49	92 221.54	3 627 361.09	10 005.53	168 268.54	1 376 246.22	1 750.92	63 342.74	157 744.60
DB1	11.24	3 100.38	38 989.17	650 748.07	2 236.83	66 416.24	192 806.29	433.70	21 276.93	28 260.58
DBK	40.49	11 003.10	119 713.20	5 723 773.42	12 085.97	211 213.81	2 336 862.62	2 819.78	105 425.97	443 193.01
DPW	21.85	3 039.34	35 751.58	3 120 529.90	4 083.18	63 482.83	1 564 405.99	1 229.61	29 143.98	373 672.02
DTE	40.43	6 725.98	62 365.61	12 449 292.22	8 727.14	161 058.58	4 918 937.39	1 458.01	71 097.87	834 711.04
EOAN	29.26	5 587.51	64 101.83	9 228 846.25	5 407.94	85 763.46	2 751 795.26	1 610.38	45 003.34	622 245.14
FME	21.21	2 807.30	40 390.46	700 928.89	3 695.55	124 150.12	334 206.74	1 133.18	52 963.98	97 659.41
FRE	16.90	2 711.75	33 894.09	340 475.18	3 680.93	69 445.98	208 713.06	422.45	17 585.19	23 680.45
HEI	9.35	3 317.76	39 276.53	701 370.92	3 524.97	73 321.34	318 912.25	611.42	19 791.14	46 476.07
HEN3	32.19	2 676.42	41 266.30	465 748.64	3 205.38	70 077.32	243 584.49	457.59	24 517.36	36 561.23
IFX	7.17	4 376.42	44 864.96	6 605 088.00	3 946.65	97 041.22	2 250 187.47	984.75	41 728.13	582 747.08
LHA	7.13	2 953.59	36 487.69	2 585 342.06	2 811.82	51 522.81	803 350.74	761.57	19 633.18	209 178.17
LIN	32.42	2 296.01	43 861.00	381 474.56	2 803.41	62 408.79	172 591.52	1 293.37	33 669.54	68 399.42
LXS	6.00	4 184.40	43 147.40	823 462.50	3 252.21	74 075.39	231 579.72	362.42	18 749.78	21 017.78
MRK	23.63	1 392.85	24 840.28	192 377.91	1 906.78	29 967.90	95 725.38	578.02	16 777.25	19 133.98
MUV2	24.40	2 929.64	50 244.27	563 913.20	3 408.33	69 053.92	226 159.97	1 258.56	36 379.85	66 005.40
RWE	20.81	5 389.74	63 971.63	2 909 747.86	5 284.01	102 906.34	928 932.98	920.52	39 324.37	121 170.85
SAP	95.68	6 538.16	67 550.11	2 476 079.29	8 518.55	150 156.54	1 256 262.95	3 144.94	94 518.34	365 707.80
SIE	91.61	7 861.28	72 557.95	2 069 163.66	12 098.28	190 943.03	1 072 561.42	3 015.94	102 190.10	196 583.27
TKA	9.95	3 556.34	41 221.83	3 014 566.10	3 291.10	54 980.92	918 368.38	622.98	22 309.68	142 966.09
VOW3	84.29	5 059.38	61 500.74	936 956.81	5 180.32	89 913.40	321 970.15	2 306.03	57 458.02	115 518.84

Table 20. Latencies, trading costs, and tick rules used in the study.

Panel A: Latency for the two exchanges links used in the strategy				
Link	One-Way Transportation Latency (ms)	Exchange Latency (ms)	Total Latency (ms)	Total Latency Used (ms)
Chi-X / Xetra	4.150	1.100	5.250	5
Chi-X / BATS	~ 0	0.165	0.165	1
Panel B: Trading costs associated with sending orders				
Exchange	Liquidity Removal (bps)	Liquidity Providing (bps)		
BATS	0.15	0.00		
Chi-X	0.30	(0.15)		
Xetra	0.36	0.36		
Panel C: Tick size (δ) rules at Xetra				
Price Range (€)	δ (€)			
[0,10)	0.001			
$[10, 50)$	0.005			
$[50, 100)$	0.01			
$[100, \infty)$	0.05			

5. Results and analysis

5.1. Empirical lead-lag relationships

Table 21 presents the mid-quote lead-lag estimation of Chi-X/Xetra and Chi-X/BATS cross-listed stocks on our data with the discrete time grid $\mathcal{G} = \{0, 1, \ldots, 50, 55, \ldots, 100, 200, \ldots, 1000, 2000, \ldots, 15000\}$ ms.

Chi-X leads almost every DAX 30 cross-listed stock also quoted at Xetra and BATS. Exceptions are HEN3 and RWE, where no definitive lead-lag relationship exists between Chi-X/Xetra and Chi-X/BATS, respectively. An important observation is that $\widehat{\ell}^*$ (measured in milliseconds) is lower-bounded by the actual latency observed between the markets in 2013, i.e., around 4–5 milliseconds for Chi-X/Xetra and around 0–1 millisecond for Chi-X/BATS (see latencies in Section 4). This demonstrates the reliability of the Hoffman et al. [39] lag estimation. Any lead-lag movement in the lagging exchange that takes longer than latency is theoretically exploitable by a HFTer. The number of potential arbitrage opportunities is presented in the next subsection.

Table 21. Mid-quote lead-lag estimation using the Hoffman et al. [39] estimator and Huth and Abergel [41] $LLR_{X,Y}$ for the links Chi-X/Xetra and Chi-X/BATS on our data.

	Chi-X / Xetra				**Chi-X / BATS**			
Ticker	**Leader**	$LLR_{X,Y}$	$\widehat{\ell}^*$ **(ms)**	$\widehat{\rho}^{HY}_{X,Y}(\widehat{\ell}^*)$	**Leader**	$LLR_{X,Y}$	$\widehat{\ell}^*$ **(ms)**	$\widehat{\rho}^{HY}_{X,Y}(\widehat{\ell}^*)$
ADS	Chi-X	1.15	10	0.025	Chi-X	2.94	4	0.034
ALV	Chi-X	2.12	8	0.046	Chi-X	4.00	2	0.157
BAS	Chi-X	1.81	8	0.034	Chi-X	3.32	1	0.039
BAYN	Chi-X	1.93	9	0.065	Chi-X	1.36	2	0.065
BEI	Chi-X	1.07	6	0.059	Chi-X	1.64	2	0.154
BMW	Chi-X	1.21	6	0.094	Chi-X	2.83	4	0.098
CBK	Chi-X	2.21	10	0.077	Chi-X	3.36	1	0.034
CON	Chi-X	1.37	7	0.039	Chi-X	1.89	10	0.033
DAI	Chi-X	1.35	7	0.052	Chi-X	1.07	1	0.051
DB1	Chi-X	1.25	5	0.031	Chi-X	3.58	2	0.120
DBK	Chi-X	1.73	5	0.100	Chi-X	2.81	4	0.105
DPW	Chi-X	1.85	8	0.060	Chi-X	2.77	1	0.060
DTE	Chi-X	2.34	9	0.039	Chi-X	2.12	1	0.206
EOAN	Chi-X	3.98	7	0.030	Chi-X	1.31	0	0.038
FME	Chi-X	1.19	7	0.035	Chi-X	2.89	2	0.032
FRE	Chi-X	1.01	9	0.025	Chi-X	2.16	1	0.085
HEI	Chi-X	1.53	6	0.033	Chi-X	1.07	1	0.306
HEN3	-	-	-	-	Chi-X	7.26	1	0.047
IFX	Chi-X	1.26	7	0.034	Chi-X	2.38	3	0.045
LHA	Chi-X	1.29	6	0.072	Chi-X	7.76	1	0.138
LIN	Chi-X	2.20	8	0.063	Chi-X	1.93	1	0.087
LXS	Chi-X	1.12	10	0.035	Chi-X	2.49	10	0.026
MRK	Chi-X	1.48	7	0.088	Chi-X	1.80	1	0.094
MUV2	Chi-X	1.90	8	0.019	Chi-X	2.89	2	0.061
RWE	Chi-X	1.27	8	0.032	-	-	-	-
SAP	Chi-X	1.56	8	0.062	Chi-X	1.30	0	0.100
SIE	Chi-X	1.92	7	0.064	Chi-X	1.55	0	0.144
TKA	Chi-X	1.59	7	0.047	Chi-X	1.55	1	0.100
VOW3	Chi-X	1.69	8	0.021	Chi-X	1.69	3	0.044

Interestingly, the fact that Chi-X is the leader of DAX 30 stocks is a direct counterexample of some earlier papers where the home market was the main source of price discovery (Frijns et al. [24]; Grammig et al. [31]; Menkveld et al. [46]; Pascual et al. [49]). Note that these papers all work on assets with differing currencies, whereas DAX 30 stocks are all listed in a single currency. However, we do not believe the exchange rate effect (or lack thereof) is the factor explaining this difference in the origin of price discovery. Indeed, the observation that Chi-X is the leader aligns with other contributions demonstrating that the most liquid and actively traded market leads price discovery (Frijns et al. [25, 26]; Poshakwale and Theobald [51]) (see Table 19 in Section 4 for stylized facts). This is also in line with the hypothesis that the market with the lowest transaction costs will be the source of price discovery (Abhyankar

[1]; Brooks et al. [8]; Ghadhab and Hellara [28]) in the case of Chi-X/Xetra relationships (see trading costs in Section 4), which is known as the “trading cost hypothesis" introduced in Fleming et al. [21]. In the case of the Chi-X/BATS relationships, even though the liquidity-removal cost is higher at Chi-X, HFTers seem to be more active at that exchange than at BATS, probably because of the higher liquidity-providing rebates given at Chi-X. Finally, since DAX 30 stocks are multiple listed, it is also logical with the fact that foreign markets will contribute more to price discovery than the home market (Ghadhab and Hellara [28]). Thus, by being colocated at Chi-X, a HFTer should have the best chance of exploiting these lead-lag relationships in DAX 30 stocks, even if Xetra is their home exchange.

From Table 21, we can answer our first question. Indeed, the exchange that is most liquid, most actively traded, and has the highest liquidity-providing rebates will lead the high-frequency prices in the case of cross-listed stocks, even if it is not the home exchange. In our application, Chi-X is the definitive leader of DAX 30 stocks, over Xetra and BATS, for the aforementioned reasons.

5.2. Econometric model performance

We choose a lag order of $D = 10$, given that trials on the first two weeks of data show that $r_{C^X_{i-D}}$ and $r_{C^Y_{i-D}}$ are always statistically insignificant in the model for $D > 12$. The model is also losing some predictive power with $D < 10$, so setting $D = 10$ is a good middle ground. The same D is used during the entire six months and for every stock. The models are recurrently trained every five days with past data and are used out-of-sample through the next five-day period, as shown in Figure 15. "Test" sections are out-of-sample periods where live trading decisions are generated based on the predictions of the models estimated on "train" periods consisting of past days. The first two five-day periods are reserved for the first training iteration, and the first out-of-sample period is the following five days. Other training frequencies were tested, but the model’s performance did not significantly change. The lag D could also be dynamically selected at each period. But, preliminary results showed that there were only minimal (or no) additional accuracy to be gained, so

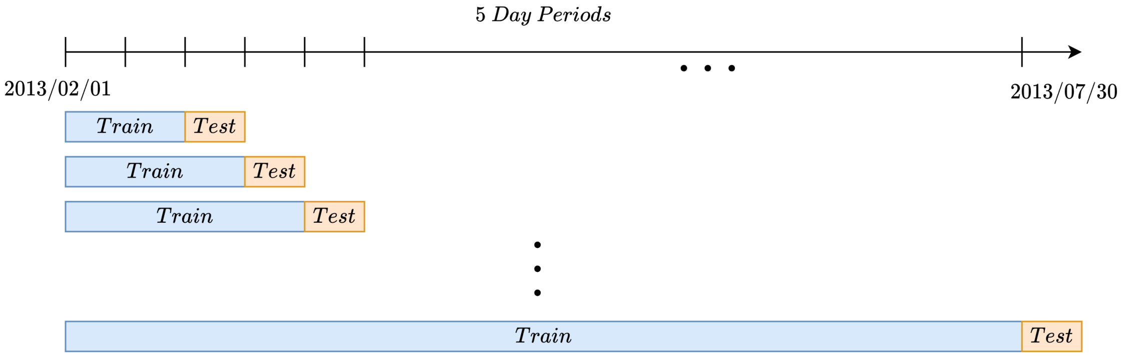

Fig. 15. Schema depicting the recurrent training and out-of-sample testing of our model every five days from February 1 to July 31, 2013

always setting $D = 10$ is an effective alternative. In fact, the difference in accuracy within $D = 10 \pm 2$ was less than ±1–2%, so D is not a main contributor to higher model accuracy.

On the other hand, the decision threshold $K \in [0, 1]$ plays an important role in selecting the right opportunities to trade on. Figure 16 exemplifies its effect on the quality of predictions and the number of potential opportunities generated by the model. Increasing K generally results in a higher accuracy in the training sample, but only up to a certain point,

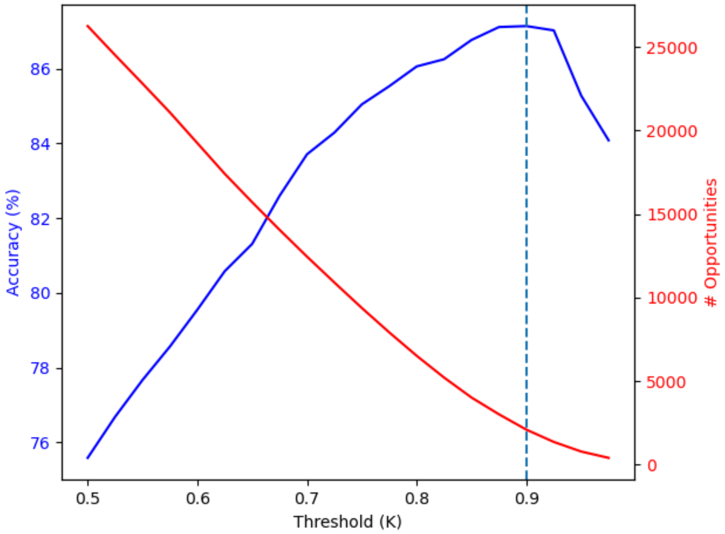

Fig. 16. Example of threshold K's effect on model performance, fitted on the bid price processes of Chi-X:DBKd and Xetra:DBK during the first training iteration. The blue line depicts the accuracy and the red one represents the number of potential opportunities, both as a function of K. The dotted vertical line is the peak of the accuracy function on the training sample.

at which it tends to decrease. It also drastically reduces the number of potential opportunities, since less and less predicted probability $\max(\widehat{\mathbf{p}}_i) \geq K$ when $K \to 1$. The peak is found on the training sample every time a model is fitted and it is used to select the trading opportunities out-of-sample. This is done independently for every stock at each exchange.

We use the model of Alsayed and McGroarty [2] as a benchmark because a clear comparison can be made between their model and ours. Moreover, the data in both studies come from similar periods. Their predictive framework currently has the best accuracy in the lead-lag arbitrage literature, so it is a suitable point of comparison. The number of potential lead-lag arbitrage opportunities on processes $(\{X_t\}, \{Y_t\})$ is defined as

$$\text{Potential Opportunities}_{\{X,Y\}} = PO_{\{X,Y\}} = \sum_{i=1}^{N} \mathbb{I}_{\left\{\widehat{R}_{C_i^Y} \neq 0\right\}},$$

which represents the number of non-null movement predictions made by a model for the next cluster of the lagging process $\{Y_t\}$. The model accuracy is then defined as

$$\text{Accuracy}_{\{X,Y\}} = \frac{1}{PO_{\{X,Y\}}} \sum_{i=1}^{N} \mathbb{I}_{\left\{\left(\widehat{R}_{C_i^Y} = R_{C_i^Y}\right) \wedge \left(\widehat{R}_{C_i^Y} \neq 0\right)\right\}},$$

the ratio of correct non-null predictions to the total number of potential opportunities. We exclude the null predictions in the accuracy measurement because they do not generate trades. We want to focus on the model's accuracy on actual opportunities. Table 22 summarizes the performance of the Alsayed and McGroarty [2] predictive model on the mid-quote from our data (see Section 3.2 for details) where δ is the tick size. For the complete per-ticker performance, see Tables 32 and 33 in Appendix B.

Table 23 presents the out-of-sample performance summary of our econometric model on the best bid and ask price processes obtained on the Chi-X/Xetra and Chi-X/BATS lead-lag relationships. Multiple dynamic thresholds are tested to study the importance of K. We begin at the peak, i.e., the values of K on the training sets that generate the highest accuracy from the set $K \in \{0.35, 0.375, 0.40, \ldots, 1\}$, and then decrease K from that starting point by increments of 0.025. For the complete per-ticker performance of our model for both best bid and ask prices processes at Xetra and BATS, see Tables 34 to 37 in Appendix B.

Table 22. Alsayed and McGroarty [2] mid-quote direction performance summary on the six months of data for multiple K^{AM}.

Threshold (K^{AM})	Xetra Accuracy (%)	Xetra Potential Opportunities	BATS Accuracy (%)	BATS Potential Opportunities	Total Accuracy (%)	Total Potential Opportunities
δ	71.7	5 187 749	70.4	4 833 712	71.1	10 021 461
2δ	70.7	1 037 573	70.9	908 307	70.8	1 945 880
3δ	66.8	351 333	68.8	285 449	67.7	636 782
4δ	64.4	192 933	67.5	148 695	65.7	341 628
5δ	63.2	126 730	66.7	95 555	64.7	222 285
6δ	62.6	85 101	66.4	63 081	64.2	148 182
7δ	62.4	57 869	66.3	42 805	64.0	100 674
8δ	62.5	38 922	65.9	28 837	64.0	67 759
9δ	62.4	26 356	66.1	19 655	64.0	46 011
10δ	62.9	18 599	65.9	14 116	64.2	32 715

Table 23. ADLMLR out-of-sample performance summary on the six months of data for multiple K.

Threshold (K)	Xetra Accuracy (%)	Xetra Potential Opportunities	BATS Accuracy (%)	BATS Potential Opportunities	Total Accuracy (%)	Total Potential Opportunities
Peak	84.2	915 666	78.3	708 580	81.6	1 624 246
Peak - 0.025	84.3	1 093 229	78.5	868 951	81.7	1 962 180
Peak - 0.050	83.9	1 262 096	78.4	1 042 930	81.4	2 305 026
Peak - 0.075	83.5	1 428 723	78.1	1 231 914	81.0	2 660 637
Peak - 0.100	82.8	1 614 729	77.6	1 401 910	80.4	3 016 639
Peak - 0.125	82.1	1 817 528	77.0	1 568 967	79.7	3 386 495
Peak - 0.150	81.3	2 028 380	76.3	1 709 488	79.0	3 737 868
Peak - 0.175	80.6	2 162 587	75.4	1 836 598	78.2	3 999 185
Peak - 0.200	79.8	2 264 035	74.6	1 878 981	77.5	4 143 016

From Tables 22 and 23, we can see that we compare favorably in terms of accuracy. As mentioned earlier, depending only on the latest cluster observation of the leading asset can be hazardous whenever the lead-lag relationship is not as strong as the ones observed in Alsayed and McGroarty [2], as defined by the $LLR_{X,Y}$. In our cross-listed stock case, fully utilizing the leading and lagging assets' past prices resulted in an average absolute increase of 10% in total accuracy. As expected, by decreasing the threshold K, we are able to increase the number of potential opportunities at the expense of a lower model accuracy. The financial effect of K is presented in the next subsection.

We also compare the performance of the ADLMLR model to a standard autoregressive distributed lag (ADL) model, where ADLMLR is a classification model trained with maximum likelihood and ADL is a closely related regression model fitted using the OLS method. In Section 3.2, we made the case that ADLMLR has a greater profitability potential than its regression counterpart, which we show here. First, we define the ADL model closest to ADLMLR:

$$r_{C_i^Y} = \alpha + \sum_{j=0}^{D-1} \beta_j r_{C_{i-j}^X} + \sum_{j=1}^{D} \gamma_j r_{C_{i-j}^Y} + \varepsilon_j$$

where $\varepsilon_j \overset{iid}{\sim} N(0,\sigma^2)$ and $D \in \mathbb{N}^+$. In order for that model's performance to be compared to ADLMLR's, the predicted directions of the total variation in cluster C_i^Y are computed as follows:

$$\widehat{R}_{C_i^Y}^{ADL} = \begin{cases} +1, & \text{if } \widehat{r}_{C_i^Y} \geq K^{ADL} \\ 0, & \text{if } -K^{ADL} < \widehat{r}_{C_i^Y} < K^{ADL} \\ -1, & \text{if } \widehat{r}_{C_i^Y} \leq -K^{ADL}. \end{cases}$$

Again, $K^{ADL} \in \mathbb{R}_0^+$ is a preset threshold found dynamically, as described at the beginning of this subsection. Notice that, when we set $D = 1$, $\widehat{\alpha} = 0$, $\widehat{\beta}_0 = 1$, $\widehat{\gamma}_1 = 0$, the model is almost equivalent to Alsayed and McGroarty [2] (they use the minimum and maximum returns within the leader's cluster, not its total return). Also, when $D = p$, $K^{ADL} = 0$, $\widehat{\alpha} = 0$, and $\widehat{\gamma}_j = 0 \ \forall j$, we get a model similar to Huth and Abergel [41], but on a quote basis instead of a trade basis. Hence, the ADL model in conjunction with the direction prediction method is a generalization of the predictive framework employed in both studies. Table 24 presents the out-of-sample performance summary of that framework on the best bid and ask price processes selected from a grid of $K^{ADL} \in \{0, \delta, 2\delta, \ldots, 10\delta\}$ with $D = 10$. At its peak, the comparable ADL model achieves an accuracy of 79.6% on a total of 1.4 million potential arbitrage opportunities. On the other hand, as seen in Table 23, the ADLMLR model can reach the same level of accuracy, but on 3.4 million arbitrage opportunities, which is over 140% more than what ADL generates. At its peak, ADLMLR's

Table 24. ADL out-of-sample performance summary on the six months of data

Threshold (K^{ADL})	**Xetra Accuracy (%)**	**Xetra Potential Opportunities**	**BATS Accuracy (%)**	**BATS Potential Opportunities**	**Total Accuracy (%)**	**Total Potential Opportunities**
Peak	84.4	634 435	75.7	777 847	79.6	1 412 282

accuracy outperforms ADL's by an absolute 2% while creating over 200,000 more potential opportunities. This demonstrates that the classification framework of ADLMLR indeed produces a greater profitability potential, as compared to an equivalent regression framework.

To understand how the leading exchange affects the predictive model's performance, we set $\beta_{j,-1} = \beta_{j,0} = 0,\ \forall j = 0, \ldots, D-1$ in the ADLMLR model so that only past cluster returns in the lagging exchange are used to generate predictions for the cross-listed stock at that same exchange. Table 25 shows the results when $K \in \{0.35, 0.375, 0.40, \ldots, 1\}$ is dynamically set at the peak. Not utilizing the lead-lag relationship between Chi-X and the lagging exchanges Xetra and BATS dramatically lowers the model's accuracy compared to Table 23. In fact, it does not significantly outperform a naive forecasting model randomly predicting positive or negative returns in the lagging exchange. This random model is able to get an accuracy of 40.1% at Xetra and 40.5% at BATS. Hence, relying only on Xetra and BATS to predict their own future returns is hardly possible because of the poor accuracy. This is in line with the efficient market hypothesis (Fama [20]). But, using prices observed at Chi-X enables accurate return predictions at lagging exchanges. This is a direct violation of the hypothesis. This is another proof of an existing lead-lag relationship for DAX 30 stocks at these three European exchanges. Additionally, when we set $\gamma_{j,-1} = \gamma_{j,0} = 0,\ \forall j = 1, \ldots, D$ without constraining the βs, ADLMLR's accuracy decreases slightly compared to Table 23. This means that the best model employs both the leading and lagging exchange prices to

Table 25. ADLMLR out-of-sample performance summary on the six months of data without the leading exchange observations ($\beta_{j,-1} = \beta_{j,0} = 0,\ \forall j = 0, \ldots, D-1$)

Threshold (K)	**Xetra Accuracy (%)**	**Xetra Potential Opportunities**	**BATS Accuracy (%)**	**BATS Potential Opportunities**	**Total Accuracy (%)**	**Total Potential Opportunities**
Peak	43.6	1 690 915	42.6	1 014 306	43.2	2 705 221

generate its predictions; this is the one used through the remainder of the paper. Huth and Abergel [41] and Alsayed and McGroarty [2] only incorporate a subset of that information, but we are able to utilize it all.

We are interested in ADLMLR's performance through time in order to make sure that it is long-lasting and well founded. Figure 17 illustrates the out-of-sample aggregated accuracy of our econometric models when K is set at peak training accuracy and $D = 10$ for every stock and every trading period. The models' out-of-sample accuracies are fairly stationary in time, varying by about 3%, and centered at the temporal mean during the entirety of our data sample. Therefore, ADLMLR is able to generate a robust predictive function based on the lead-lag effect observed between Chi-X/Xetra and Chi-X/BATS. The model performs on average 6% better at Xetra and it constantly outperforms the one fitted at BATS.

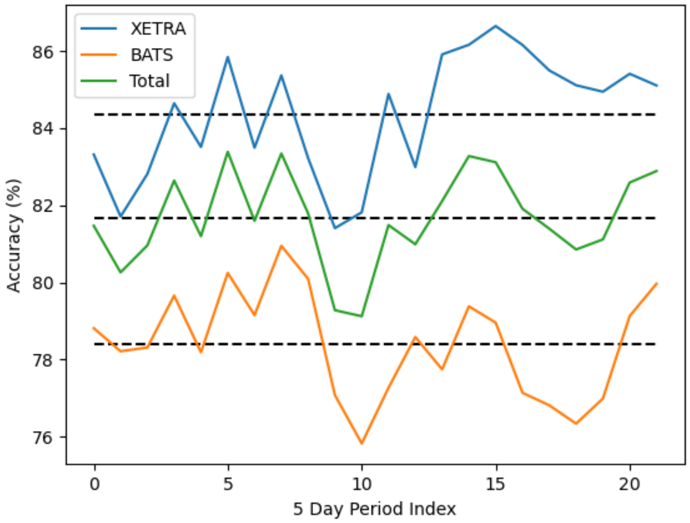

Fig. 17. Out-of-sample accuracy in time, weighted on Table 21 selected DAX 30 stocks of our econometric models for Xetra and BATS at each 5-day period from February 1 to July 31, 2013.

From Table 23 and Figure 17, we demonstrate that if there is a lead-lag relationship between any two assets, an adequate econometric model fully utilizing current and past observations of both assets is able to predict the lagging returns with respectable accuracy. In our case, a generalized form of autoregressive logistic regression can predict the next cluster movement of Xetra's and BATS' best bid and ask prices out-of-sample with an average accuracy exceeding 80%. This is possible because Chi-X led the DAX 30 cross-listed stocks prices.

5.3. Statistical arbitrage performance

We compute the performance of the HFT arbitrage strategy of Section 3.3 in two scenarios to determine the lead-lag relationships' financial significance. In the first scenario, we only consider the first half of latency. We observe the LOBs of Xetra and BATS at a delay because the physical distance between these exchanges and Chi-X causes the information to arrive late at that location. In the second scenario, the first half of latency is still considered, but now orders sent to Xetra and BATS also arrive at a delay to account for the second half of latency. Both scenarios consider trading costs and assume the colocation of a server at Chi-X. This allows us to empirically study the effect of latency on the arbitrage strategy's performance.

Table 26 details the performance of the HFT strategy when latency is considered in the case of information arrival, but not when sending orders (scenario 1). By being colocated at Chi-X, we receive Xetra's TAQ data five milliseconds after it is sent by the exchange, and BATS' data is received after one millisecond. But, orders are immediately integrated into Xetra's and BATS's LOBs whenever they are sent by the strategy. As in Table 23, we begin at the peak, i.e., the values of K on the training sets that generate the highest accuracy from the set $K \in \{0.35, 0.375, 0.40, \ldots, 1\}$, and then decrease K from that starting point by increments of 0.025.

We stop at $K = \text{Peak} - 0.200$ because it is the point at which the strategy's profitability starts to diminish and continues to do so past that threshold. Table 27 presents the performance of the HFT strategy when latency is also included when sending orders to the market, while still considering information arrival latency (scenario 2), meaning that orders sent to Xetra take five milliseconds to arrive in the LOB, and orders sent to BATS arrive after one millisecond from a colocated server at Chi-X. Full latency is thus considered, being the most realistic scenario, in accounting for important market frictions.

Table 26. Performance summary of the HFT arbitrage strategy on six months of 2013 data for the first scenario and multiple Ks, all in €.

Threshold	**Xetra Profits**	**Xetra**	**BATS Profits**	**BATS**	**Total Profits**	**Total Net Profits**
(K)	**(before costs)**	**Net Profits**	**(before costs)**	**Net Profits**	**(before costs)**	
Peak	607 013	121 976	405 893	327 640	1 012 906	449 616
Peak - 0.025	739 347	137 269	521 868	422 887	1 261 216	560 155
Peak - 0.050	880 287	151 088	634 577	514 435	1 514 864	665 523
Peak - 0.075	1 043 393	173 071	730 251	589 172	1 773 644	762 243
Peak - 0.100	1 236 158	198 566	800 106	642 167	2 036 264	840 733
Peak - 0.125	1 443 021	214 671	847 086	674 313	2 290 107	888 984
Peak - 0.150	1 667 882	246 440	874 900	691 163	2 542 782	937 603
Peak - 0.175	1 878 797	290 017	874 617	681 707	2 753 413	971 725
Peak - 0.200	2 058 342	318 596	849 585	652 701	2 907 927	971 296

Table 27. Performance summary of the HFT arbitrage strategy on six months of 2013 data for the second scenario and multiple Ks, all in €.

Threshold	**Xetra Profits**	**Xetra**	**BATS Profits**	**BATS**	**Total Profits**	**Total Net Profits**
(K)	**(before costs)**	**Net Profits**	**(before costs)**	**Net Profits**	**(before costs)**	
Peak	555 629	99 891	423 990	346 240	979 618	446 131
Peak - 0.025	678 371	111 283	542 332	443 910	1 220 703	555 193
Peak - 0.050	809 847	120 902	657 113	537 573	1 466 961	658 475
Peak - 0.075	962 084	136 661	752 228	614 798	1 714 312	751 459
Peak - 0.100	1 146 414	158 360	828 672	671 389	1 975 086	829 749
Peak - 0.125	1 349 241	174 914	879 733	707 632	2 228 974	882 546
Peak - 0.150	1 566 425	203 051	908 245	725 161	2 474 670	928 212
Peak - 0.175	1 773 586	244 850	910 667	718 406	2 684 254	963 257
Peak - 0.200	1 945 885	268 123	885 588	689 349	2 831 473	957 471

Comparing Table 26 with Table 27, we notice that adding latency to the orders sent by the HFT strategy plays an important role in its net profitability, especially at Xetra. Indeed, net profits at that exchange are reduced by 15%—20%, but the strategy still remains profitable. On the other hand, net profits at BATS do not change dramatically (around 5% change). The geographical proximity of BATS to Chi-X and its lower trading activity and liquidity compared to Xetra makes it so that latency does not play an important role on the net profitability. Because of its higher trading costs, its geographical distance to the leading exchange, and its higher level of trading and quoting activity, as compared to BATS, generating net profits from lead-lag arbitrage at Xetra is more challenging. From these results, we show that a HFTer is able to exploit the lead-lag relationship that exists for most DAX 30 stocks cross-listed at Xetra, Chi-X, and BATS even when full latency, non-execution risk, and trade costs are considered. From Table 27, we see that a HFTer can realistically generate an annual net profit of over €1.9 million on DAX 30 stocks alone from the three exchanges, or more than €33,000 on average per cross-listed stock, per exchange. Table 28 presents the detailed performance of the Alsayed and McGroarty [2] strategy with the most accurate K^{AM}.

Table 28. Detailed performance of the Alsayed and McGroarty [2] strategy on six months of 2013 data in the second scenario with the most accurate threshold $K^{AM} = \delta$.

Exchange	**Xetra**	**BATS**
Gross Profit(€)	29 647	2 415
Loss (€)	-11 530 611	-23 819 384
Trading Costs (€)	-1 597 281	-317 594
Total Net Profit (€)	-13 098 246	-24 134 563
Median Net Daily Profit (€)	-110 407	-201 945
Mean Net Daily Profit (€)	- 115 914	-213 580
Most Profitable Date (Net Profit, €)	5/20/2013 (-59 006)	7/23/2013 (-97 640)
Fifth Most Profitable Date (Net Profit, €)	7/22/2013 (-63 408)	7/22/2013 (-112 638)
Least Profitable Date (Net Profit, €)	2/26/2013 (-290 537)	5/2/2013 (-448 108)
Fifth Least Profitable Date (Net Profit, €)	2/21/2013 (-136 762)	2/21/2013 (-239 112)
Median Trade Time (s)	0.050	0.021
Mean Trade Time (s)	2.17	2.17
# Net Profitable Trades	27 350	1 917
# Net Unprofitable Trades	4 196 171	4 124 243
# Trades	4 223 521	4 126 160
Net Profitable Trades (%)	0.65	0.05
Mean Volume per Trade	100	100
Mean Net Profit per Profitable Trade (€)	0.68	1.17
Mean Net Profit per Unprofitable Trade (€)	-3.12	-5.85

The most accurate version of the Alsayed and McGroarty [2] mid-quote strategy is not able to cover the bid-ask spread and the transaction costs. Almost 100% of the trades in this strategy are not profitable because a variation in the best bid (when closing a long position) and in the best ask (when closing a short position) greater than the bid-ask spread, plus the transaction costs, is necessary within a single cluster, which lasts on average around two seconds at both exchanges. This profitable situation occurs 0.65% of the time at Xetra and 0.05% at BATS. Larger values of K^{AM} do not generate better results in terms of net profit per trade, and no K^{AM}s generate a net profitable strategy.

We also demonstrate that a mid-quote–based market order HFT strategy, like the one in Huth and Abergel [41] and Alsayed and McGroarty [2] is not viable in practice. To do so, we assume a perfect model that is always able to predict the exact mid-quote return of the lagging asset's next cluster. If that return is above (under) a threshold $K^{Perfect}$ ($-K^{Perfect}$), the strategy opens a long (short) position with a buy (sell) market order at the best ask (bid) right before the lagging asset's next cluster begins. The position is then closed with an opposite market order precisely when the lagging asset's cluster ends. This is the buy-and-hold HFT strategy of Alsayed and McGroarty [2]. Huth and Abergel [41] employ the same type of strategy, but on a trade basis with a threshold of 0. Table 29 presents this best case mid-quote–based market order HFT strategy on our data in the second scenario.

Table 29. Performance of a best case mid-quote–based market order HFT strategy on six months of 2013 data in the second scenario for multiple $K^{Perfect}$.

Threshold	**# Trades**	**Net Profitable**	**Gross Profit**	**Loss**	**Trading Costs**	**Total Net Profits**
$(K^{Perfect})$		**Trades (%)**	**(€)**	**(€)**	**(€)**	**(€)**
δ	11 383 116	0.50	69 459	-44 766 290	-2 677 747	-47 374 578
2δ	2 881 086	1.36	49 936	-17 000 432	-596 335	-17 546 832
3δ	1 226 077	1.67	30 536	-10 368 212	-197 801	-10 535 476
4δ	723 858	1.40	20 858	-7 427 307	-94 246	-7 500 695
5δ	427 531	1.30	15 601	-5 414 912	-52 933	-5 454 244
6δ	303 438	1.27	13 100	-4 097 537	-34 348	-4 118 785
7δ	180 751	1.50	11 385	-2 990 559	-21 216	-3 000 390
8δ	113 714	1.82	10 196	-2 332 528	-14 061	-2 336 393
9δ	71 894	2.15	9 057	-1 873 618	-9 477	-1 874 038
10δ	47 844	2.54	8 251	-1 537 994	-6 420	-1 535 892

Even though the predictive model is perfectly accurate on the next mid-quote return of the lagging asset, gross profits never cover the bid-ask spread cost of market orders. This is the only source of losses in Table 29. Thus, it is impossible to profit from high-frequency lead-lag arbitrage from mid-quote return predictions and a market order–based HFT strategy. It also shows that at the millisecond scale, asset returns rarely cover market order trading costs. This means that the trading signal of Stübinger [56] would also generate inconsiderable profits in this setting. Switching from market orders to limit orders eliminates the necessity of covering the bid-ask spread and facilitates access to profitability. It is also important to know what side(s) of the LOB will generate the non-zero mid-quote return to capture arbitrage opportunities and mid-quote returns do not provide that information. Predicting the best bid and best ask returns allows better-informed trading decisions. Table 30 presents the detailed performance of our limit order–based strategy with the most profitable K in the second scenario.

Table 30. Detailed performance of the HFT strategy on six months of 2013 data in the second scenario with the most profitable threshold $K = \text{Peak} - 0.175$.

Exchange	**Xetra**	**BATS**
Gross Profit (€)	3 365 103	2 108 945
Loss (€)	-1 591 517	-1 198 278
Trading Costs (€)	-1 528 736	-192 261
Total Net Profit (€)	244 850	718 406
Median Net Daily Profit (€)	1 943	6 072
Mean Net Daily Profit (€)	2 167	6 358
Most Profitable Date (Net Profit, €)	6/11/2013 (9 987)	2/26/2013 (16 118)
Fifth Most Profitable Date (Net Profit, €)	6/24/2013 (5 723)	2/25/2013 (11 053)
Least Profitable Date (Net Profit, €)	5/2/2013 (-2 290)	5/9/2013 (2 812)
Fifth Least Profitable Date (Net Profit, €)	2/21/2013 (1 237)	7/26/2013 (4 173)
Median Trade Time (s)	1.02	1.44
Mean Trade Time (s)	27.82	28.45
# Net Profitable Trades	1 158 049	1 002 859
# Net Unprofitable Trades	223 452	223 998
# Trades	1 381 501	1 226 857
Net Profitable Trades (%)	83.83	81.74
Mean Volume per Trade	503.64	352.29
Mean Net Profit per Profitable Trade (€)	1.79	1.95
Mean Net Profit per Unprofitable Trade (€)	-8.20	-5.51

Gross profits are considerable in both exchanges. But, losses incurred from execution-related risks are also sizeable, drastically decreasing the net profitability of the strategy, by approximately 50%. Losses occur whenever the model predictions are wrong, which directly results in limit orders not being executed because the lagging assets' level 1 prices have drifted away from the specified limit price. At that point, losses are cut by stop limit orders. When these limit orders are also not executed, market orders are sent to finally close the position after M minutes (15 minutes for Xetra and 20 for BATS; see Appendix A for details). Losses can also occur even when the model is right, but limit orders remain in the queue without ever being executed.

Exchange trading costs are also significant, especially at Xetra, given its prohibitive fee structure relative to BATS. This was expected, given the large number of trades and their limited profitability because of the brief holding period typical of HFT strategies. Overall, considering losses and trading costs, a net profit margin of 7% was obtained at Xetra and 34% at BATS, where the significant difference stems from that difference in their fee structure and from the longer latency to trade on Xetra from Chi-X. All order types are expensive at

Xetra, whereas liquidity-providing orders are free at BATS and liquidity-taking fees are less than half of Xetra's (see Table 20 for all fees).

Median trading times are quick at both exchanges, though slightly longer at BATS because of its lower level of trading activity compared to Xetra (see Table 19). Mean trading times are greater than the median, given the non-execution risk of limit orders, which can stay for up to M minutes in the LOB without being executed. The proportions of net profitable trades are in line with model accuracy for both exchanges. Once again, we notice the importance of execution-related risks from the difference between the performance of profitable and non-profitable trades. In fact, the mean loss incurred is over 4.58 times greater than the mean profit per trade at Xetra, and the same ratio is over 2.82 at BATS. Without risk management procedures, these ratios are even greater. Table 31 presents the detailed performance of the HFT strategy excluding stop-limit orders, maximum level 1 price variation, and the no-microstructure-change rule (see Appendix A for details). We leave the time breaker of M minutes before closing positions; otherwise they can stay open for days and no trade occurs in that time because the strategy waits for the previous position to close before opening the next. This is a consequence of the non-execution risk of limit orders.

As expected, the ratio of mean loss to mean profit per trade incurred at Xetra climbs from 4.58 to 6.70 and soars from 2.82 to 18.50 at BATS. More importantly, the net profitability decreased by 27% at Xetra and by 39% at BATS. Nonetheless, the strategy remains profitable at both exchanges. The largest difference between Table 30 and Table 31 comes from the absence of stop-limit orders. Without them, the positions stay open as long as the profit-taking limit orders are not executed, up to M minutes. The average trade duration more than doubles, hence reducing the number of arbitrage opportunities captured by about the same quotient. Risk management procedures are thus useful in preventing large losses by mitigating the non-execution risk of limit orders while also closing positions rapidly when prices drift away for them.

Figure 18 presents the cumulative net profit of the most profitable version of the strategy in scenario 2 (see Table 30). The strategy has minimal drawdown and constantly generates

Table 31. Detailed performance of the HFT strategy on six months of 2013 data in the second scenario with the most profitable threshold K = Peak − 0.175 without risk management.

Exchange	Xetra	BATS
Gross Profit (€)	1 569 778	1 146 940
Loss (€)	-767 203	-604 726
Trading Costs (€)	-624 995	-100 639
Total Net Profit (€)	177 580	441 575
Median Net Daily Profit (€)	1 454	3 767
Mean Net Daily Profit (€)	1 572	3 908
Most Profitable Data (Net Profit, €)	6/13/2013 (4 708)	3/6/2013 (15 650)
Fifth Most Profitable Date (Net Profit, €)	5/23/2013 (3 910)	2/27/2013 (7 581)
Least Profitable Date (Net Profit, €)	7/5/2013 (-2 871)	7/19/2013 (-2 684)
Fifth Least Profitable Date (Net Profit, €)	2/21/2013 (1 321)	6/21/2013 (699)
Median Trade Time (s)	1.80	1.76
Mean Trade Time (s)	75.22	79.81
# Net Profitable Trades	456 981	546 489
# Net Unprofitable Trades	55 701	17 172
# Trades	512 682	563 661
Net Profitable Trades (%)	89.14	96.95
Mean Volume per Trade (€)	338.45	213.47
Mean Net Profit per Profitable Trade (€)	2.14	1.92
Mean Net Profit per Unprofitable Trade (€)	-14.33	-35.52

a positive net profit on a daily basis. Table 30 answers our final question. If a lead-lag relationship exists between two assets and if a predictive model is able to exploit it, a HFTer can in fact realistically earn a profit. As shown in the same table, the execution-related

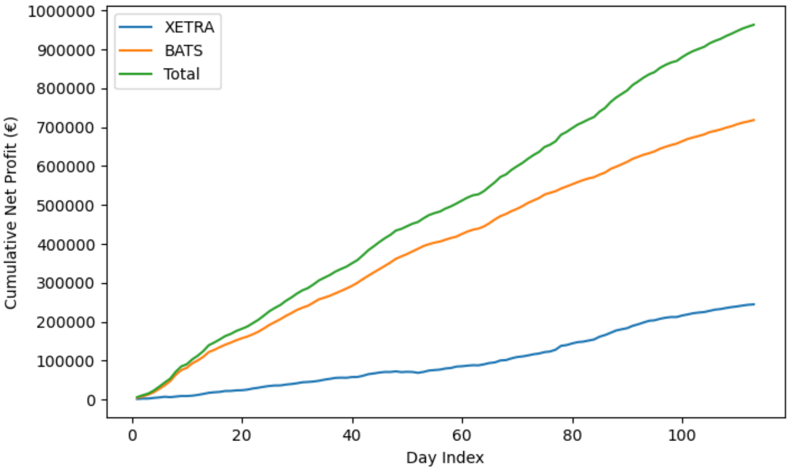

Fig. 18. Cumulative net profit of the HFT strategy on a daily basis for Table 21 selected DAX 30 stocks from February 1 to July 31 2013 in the second scenario with the most profitable threshold $K = Peak - 0.175$.

risks were the main impediment to lead-lag arbitrage, followed by trading costs and latency, based on Table 26 versus Table 27. Nonetheless, an HFT strategy that was colocated at the leading exchange and that relied mainly on limit orders was able to profit from the lead-lag relationship that existed between DAX 30 stocks cross-listed at Xetra, Chi-X, and BATS in 2013. But in practice, additional costs arise when deploying such a strategy. As detailed in this section, lead-lag arbitrage is not completely risk free, and constant monitoring is necessary to maintain profitability in the long term. Indeed, structural changes in any exchange, e.g., trading fees or rules, can impact the lead-lag relationships, and the model must be regularly trained (performance overseen, etc.), thus requiring human intervention. There are also substantial costs related to accessing colocation and data services at multiple exchanges, which reduces further the profitability presented in this section. The strategy then mostly becomes interesting for sophisticated HFTers that are already paying for these resources in their general activities.[27]

6. Conclusion

In this paper, we investigate the existence, predictability, and profitability of lead-lag relationships at a high frequency with an application to DAX 30 stocks cross-listed at Xetra in Frankfurt, and Chi-X and BATS, both in London, during six months of 2013. Using the robust lead-lag estimator of Hoffman et al. [39] and the lead-lag ratio of Huth and Abergel [41], we first show that Chi-X leads level 1 prices by mere milliseconds. This is in line with previous studies showing that the most actively traded, liquid, and least expensive exchange will ultimately be the price discovery origin of arbitrage-linked assets. The lead-lag estimation demonstrates the great interconnectedness between the three exchanges by showing that their lag is approaching, or even equating, the physical speed limit at which information could travel between them at that time. This level of precision is the highest in the cross-listed stocks lead-lag literature.

[27] For example, Deutsche Börse Group offers colocation services starting at around €5,000/month. See https://www.xetra.com/xetra-en/technology/co-location-services for detailed pricing (accessed April 6, 2023).

After establishing the existence of lead-lag relationships in DAX 30 cross-listed stocks, we develop a new predictive modeling framework based on the concept of clusters proposed by Alsayed and McGroarty [2], in conjunction with a new, generalized version of the autoregressive logistic regression. Clusters allow us to depart from uniformly sampled observations to instead employ the unadulterated LOB updates. Our econometric model employs past and current asset prices to forecast a classification of the next clusters' return: positive, null, or negative. This probabilistic framework generates an out-of-sample return accuracy exceeding 80%, with a solidly maintained performance throughout our data period, thereby comparing advantageously to the other models put forth in the literature. Indeed, the proposed approach is able to detect substantially more potential arbitrage opportunities, with an even greater accuracy than previous regression models.

We then introduce a new high-frequency trading strategy built around our predictive model to profit from the detected lead-lag relationships. Previous studies failed to generate viable high-frequency strategies because of the steep costs associated with market orders (Brooks et al. [8]; Huth and Abergel [41]). In these studies, paying the bid-ask spread and the exchange trading costs was too prohibitive to exploit intraday lead-lag relationships. To go further, we empirically demonstrate the non-viability of mid-quote and market order–based strategies in the context of high-frequency lead-lag arbitrage. The results show the quasi-impossibility of such a strategy to cover even the bid-ask spread when lags exist at the sub-second scale. The strategy we propose relies instead on limit orders and LOB signals to cut on these costs, at the expense of adding a non-execution risk. In a scenario where major market frictions are present, we demonstrate that high-frequency traders could realistically earn a profit with our limit order–based strategy. More precisely, they could generate an annual net profit above €1.9 million from DAX 30 stocks alone and only two exchanges (Xetra and BATS) with a colocated server at Chi-X. We show that execution-related risks, trading costs, and latency (in that order) are important impediments to lead-lag arbitrage, and that risk management measures are necessary to alleviate their impact on profitability. Also, important human capital and technological costs are associated to deploying and monitoring the strategy.

Our goal was to demonstrate how a high-frequency trader would theoretically be able to profit from lead-lag arbitrage and empirically show that possibility with a pragmatic approach. We intended to develop a complete framework incorporating the detection, prediction, and trading of lead-lag relationships for any pair of assets. The framework empirically achieved that for cross-listed stocks, hence advancing knowledge on lead-lag in high-frequency markets and answering queries about their financial importance (Basnarkov et al. [3]; Curme et al. [14]). The proposed framework is also general enough to be used on any type of assets. However, detecting and exploiting lead-lag relationships in assets listed in a single currency is arguably easier since there is no exchange rate to consider. Indeed, one would have to add the effect of this third process in the lead-lag relationship. In that sense, new research exploring the effects of exchange rates in lead-lag relationships in LOB data would be important to expand the proposed framework to a greater set of international assets. Furthermore, as shown in Russell and Engle [54] for a single asset, the duration between past market events can be an additional source of predictability in price changes. It would be interesting to study the time dynamics between the leading and lagging asset clusters, which could then be potentially exploited by the prediction model and the HFT strategy.

Our study covered the application of high-frequency lead-lag relationships in an arbitrage context. Li et al. [32] demonstrate how the daily lead-lag effect significantly improves the profitability of alpha-factor strategies. In that sense, the statistical relationship, predictive model, and backtesting methodology presented in this paper could also be investigated for other types of strategies, like market making. Being able to predict an asset's level 1 prices from another related and leading asset would probably prove beneficial for market markers. It would also be worthwhile to quantify the financial viability of lead-lag relationships in other asset classes and markets, and during different time periods with the proposed framework, or any other that might come.

Acknowledgments

Financial support from the Canada Research Chair in Risk Management, Canada First Research Excellence Fund, and IVADO is graciously acknowledged.

References

[1] Abhyankar, A. (1995). Return and volatility dynamics in the FT-SE 100 stock index and stock index futures markets. *Journal of Futures Markets*, 15(4):457–488.

[2] Alsayed, H. and McGroarty, F. (2014). Ultra-high-frequency algorithmic arbitrage across international index futures. *Journal of Forecasting*, 33:391–408.

[3] Basnarkov, L., Stojkoski, V., Utkovski, Z., and Kocarev, L. (2020). Lead-lag relationships in foreign exchange markets. *Physica A: Statistical Mechanics and its Applications*, 539(1).

[4] Bilodeau, Y. (2013). Xetra parser [computer software]. HEC Montréal.

[5] Bollen, N., O'Neill, M., and Whaley, R. (2017). Tail wags dog: Intraday price discovery in VIX markets. *Journal of Financial Markets*, 37(5):431–451.

[6] Bonney, G. (1987). Logistic regression for dependent binary observations. *Biometrics*, 43(4):951–973.

[7] Brooks, C., Garrett, I., and Hinich, M. (1999). An alternative approach to investigating lead-lag relationships between stock and stock index futures markets. *Applied Financial Economics*, 9:605–613.

[8] Brooks, C., Rew, A., and Ritson, S. (2001). A trading strategy based on the lead-lag relationship between the spot index and futures contract for the FTSE 100. *International Journal of Forecasting*, 17:31–44.

[9] Broyden, C. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6:76–90.

[10] Budish, E., Cramton, P., and Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quaterly Journal of Economics*, 130(4):1547–1621.

[11] Chan, K. (1992). A further analysis of the lead-lag relationship between the cash market and stock index futures market. *The Review of Financial Studies*, 5(1):123–152.

[12] Chen, Y., Da, Z., and Huang, D. (2019). Arbitrage trading: The long and the short of it. *The Review of Financial Studies*, 32:1608–1646.

[13] Chen, Y. and Gau, Y. (2010). News announcements and price discovery in foreign exchange spot and futures markets. *Journal of Banking and Finance*, 34:1628–1636.

[14] Curme, C., Tumminello, M., Mantegna, R., Stanley, H., and Kenett, D. (2015). Emergence of statistically validated financial intraday lead-lag relationships. *Quantitative Finance*, 15(8):1375–1386.

[15] Deutsche Börse Group (2013). Presentation: Investor day 2013. Available at https://www.deutsche-boerse.com/dbg-en/investor-relations/presentations. Accessed on 7 September 2022.

[16] Deutsche Börse Group (2012). 124/2012 amendment to the price list for the utilization of the exchange edp of fwb frankfurt stock exchange and of the edp xontro. Available at https://www.deutsche-boerse-cash-market.com/dbcm-en/newsroom/circulars/Xetra-circulars-mailings. Accessed on 7 September 2022.

[17] Dimpfl, T. and Jung, R. (2012). Financial market spillovers around the globe. *Applied Financial Economics*, 22(1):45–57.

[18] Engle, R. and Granger, C. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica*, 55:251–276.

[19] Epps, T. (1979). Comovements in stock prices in the very short-run. *Journal of the American Statistical Association*, 74(366):291–298.

[20] Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

[21] Fleming, J., Ostdiek, B., and Whaley, R. (1996). Trading costs and the relative rates of price discovery in stock, futures and options markets. *Journal of Futures Markets*, 4:353–387.

[22] Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, 13(3):317–322.

[23] Foucault, T. and Biais, B. (2014). HFT and market quality. *Bankers, Markets & Investors*, 128:5–19.

[24] Frijns, B., Gilbert, A., and Tourani-Rad, A. (2010). The dynamics of price discovery for cross-listed shares: Evidence from Australia and New Zealand. *Journal of Banking and Finance*, 34:498–508.

[25] Frijns, B., Gilbert, A., and Tourani-Rad, A. (2015). The determinants of price discovery: Evidence from US-Canadian cross-listed shares. *Journal of Banking and Finance*, 59:457–468.

[26] Frijns, B., Indriawan, I., and Tourani-Rad, A. (2018). The interactions between price discovery, liquidity and algorithmic trading for U.S.-Canadian cross-listed shares. *International Review of Financial Analysis*, 56:136–152.

[27] Frino, A. and West, A. (2003). The impact of transaction costs on price discovery: Evidence from cross-listed stock index futures contracts. *Pacific-Basin Finance Journal*, 11:139–151.

[28] Ghadhab, I. and Hellara, S. (2016). Price discovery of cross-listed firms. *International Review of Financial Analysis*, 44:177–188.

[29] Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, 24(109):23–26.

[30] Gonzalo, J. and Granger, C. (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business and Economic Statistics*, 13:1–9.

[31] Grammig, J., Melvin, M., and Schlag, C. (2005). Internationally cross-listed stock prices during overlapping trading hours: price discovery and exchange rate effects. *Journal of Empirical Finance*, 12:139–164.

[32] Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

[33] Hasbrouck, J. (1995). One security, many markets: determining the contributions to price discovery. *The Journal of Finance*, 50:1175–1199.

[34] Hasbrouck, J. (2021). Price discovery in high resolution. *Journal of Financial Econometrics*, 19(3):395–430.

[35] Hasbrouck, J. and Saar, G. (2013). Low-latency trading. *Journal of Financial Markets*, 16:646–679.

[36] Hayashi, T. and Koike, Y. (2018). Wavelet-based methods for high-frequency lead-lag analysis. *SIAM Journal of Financial Mathematics*, 9(4):1208–1248.

[37] Hayashi, T. and Yoshida, N. (2005). On covariance estimation of non-synchronously observed diffusion processes. *Bernouilli*, 11(2):359–379.

[38] Herbst, A., McCormack, J., and West, E. (1987). Investigation of a lead-lag relationship between spot stock indices and their futures contracts. *The Journal of Futures Markets*, 7(4):373–381.

[39] Hoffman, M., Rosenbaum, M., and Yoshida, N. (2013). Estimation of the lead-lag parameter from non-synchronous data. *Bernouilli*, 19(2):426–461.

[40] Hou, K. (2007). Industry diffusion and the lead-lag effect in stock returns. *The Review of Financial Studies*, 20(4):1113–1138.

[41] Huth, A. and Abergel, F. (2014). High frequency lead/lag relationships - Empirical facts. *Journal of Empirical Finance*, 26:41–58.

[42] Jong, F. and Nijman, T. (1997). High frequency analysis of lead-lag relationships between financial markets. *Journal of Empirical Finance*, 4(2–3):259–277.

[43] Judge, A. and Reancharoen, T. (2014). An empirical examination of the lead-lag relationship between spot and futures markets Evidence from Thailand. *Pacific-Basin Finance Journal*, 29:335–358.

[44] Kawaller, I., Koch, P., and Koch, T. (1987). The temporal price relationship between S&P 500 futures and the S&P 500 index. *The Journal of Finance*, 42(5):1309–1329.

[32] Li, Y., Wang, T., Sun, B., and Liu, C. (2022). Detecting the lead–lag effect in stock markets: Definition, patterns, and investment strategies. *Financial Innovation*, 8(51).

[46] Menkveld, A., Koopman, S., and Lucas, A. (2007). Modeling around-the-clock price discovery for cross-listed stocks using state space methods. *Journal of Business and Economic Statistics*, 25(2):213–225.

[47] Nguyen, V., Shaker, M., and Hüllermeier, E. (2022). How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111:89–122.

[48] O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics*, 116:257–270.

[49] Pascual, R., Pascual-Fuster, B., and Climent, F. (2006). Cross-listing, price discovery and the informativeness of the trading process. *Journal of Financial Markets*, 9:144–161.

[50] Pomponio, F. and Abergel, F. (2013). Multiple-limit trades: Empirical facts and application to lead–lag measures. *Quantitative Finance*, 13(5):783–793.

[51] Poshakwale, S. and Theobald, M. (2004). Market capitalisation, cross-correlations, the lead/lag structure and microstructure effets in the Indian stock market. *Journal of International Financial Markets, Institutions and Money*, 14(4):385–400.

[52] Poutré, C., Dionne, G., and Yergeau, G. (2021). International high-frequency arbitrage for cross-listed stocks. Available at https://ssrn.com/abstract_id=3890433.

[53] Putniņš, T. (2013). What do price discovery metrics really measure? *Journal of Empirical Finance*, 23:68–83.

[54] Russell, J. and Engle, R. (2005). A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial– autoregressive conditional duration model. *Journal of Business & Economic Statistics*, 23(2):166–180.

[55] Shanno, D. (1970). Conditioning of quasi-newton methods for function minimizations. *Mathematics of Computation*, 24(111):647–656.

[56] Stübinger, J. (2019). Statistical arbitrage with optimal causal paths on high-frequency data of the S&P 500. *Quantitative Finance*, 19(6):921–935.

[57] Tse, Y. (1995). Lead-lag relationship between spot index and futures price of the Nikkei stock average. *Journal of Forecasting*, 14:553–563.

[58] Yang, J., Yang, Z., and Zhou, Y. (2012). Intraday price discovery and volatility transmission in stock index and stock index futures markets: Evidence from China. *The Journal of Futures Markets*, 32(2):99–121.

[59] Zhang, L. (2011). Estimating covariation: Epps effect, microstructure noise. *Journal of Econometrics*, 160(1):33–47.

Appendix A. High-Frequency Strategy — Additional Details

The strategy has two important variables controlling its performance: the time breaker's delay M, and the order volume $V_t^{Bid/Ask}$. To select M at Xetra and BATS, we tested its financial effect on the first out-of-sample period. We ran the HFT strategy in that timeframe with $M \in \{5, 6, 7, 8, 9, 10, 15, 30, 60, 90, 120, 300, 600, 900, 1200, 1800, 3600\}$ seconds at the two exchanges. $M = 900$ seconds produced the greatest profitability in that first period at Xetra, and $M = 1200$ seconds at BATS. These values where then set for the entirety of our data, since dynamically selecting them (like we did for K) is computationally very expensive. As shown in Figure 18, net profits are fairly constant in time, a sign that the strategy does not suffer from a preset M.

As for $V_t^{Bid/Ask}$, it follows the median level 1 volume of the last 500 LOB updates, rounded to the closest lowest 100 to trade on round lots. Using more LOB updates does not significantly affect the volume sent by the strategy and does not greatly impact the strategy's performance. More formally, given LOB update indices $t = 1,2,\ldots,T$, the order volume that is sent by the HFT strategy is calculated by

$$V_t^{Bid/Ask} = 100 \left\lfloor \frac{\tilde{v}_t^{Bid/Ask}}{100} \right\rfloor, \forall t \geq 500$$

where $\tilde{v}_t^{Bid/Ask}$ is the empirical median of the sample $v_{t-499:t}^{Bid/Ask}$, for $v_t^{Bid/Ask} \in \mathbb{N}^+$ the best bid/ask volume at index t. No order is sent to the market before observing 500 LOB updates. This is done independently for every stock at Xetra and BATS. Using a windowed median volume limits the market impact of the strategy and the liquidity risk, because the orders dynamically and conservatively follow the liquidity present in the LOB.

To mitigate risk even more, orders are only sent when three conditions are respected:

1. The last in-cluster return of the leader $C_{i,n_i^X}^X$ is not generated by a trade;
2. The realized local variation of level 1 price at the lagging exchange is under a preset threshold;

3. No microstructure change has occurred.

The first condition is present so that the strategy does not to open a position whenever child orders hit the same ticker at multiple exchanges and at the same time. When that occurs, the LOBs of all exchanges move in the same direction at the same time. The strategy cannot profit from that situation since it depends on delayed movements of the LOB at the lagging exchange.

The second condition limits execution-related risks by not opening a position when the volatility of level 1 prices of the LOB is too great, as measured from the previous 50 prices. Given LOB update indexes $t = 1{,}2, \dots T$, the realized local variation is defined as

$$RLV_t^{Bid/Ask} = \sum_{i=0}^{49} \left| p_{t-i}^{Bid/Ask} - p_{t-i-1}^{Bid/Ask} \right|,$$

where $p_t^{Bid/Ask} \in \mathbb{R}^+$ the best bid/ask price at index t. Whenever $RLV_t^{Bid/Ask} > \delta W$ for $\delta \in \mathbb{R}^+$ the tick size and $W \in \mathbb{R}_0^+$ a preset threshold, the strategy does not send orders. W is found from the set $\{5, 10, 25, 50, 75, 100, 150, 200, 250, 500\}$ in the same way as M. $W = 100$ at Xetra and $W = 25$ at BATS.

The third condition relates to changes in the tick size of the stock. Whenever this microstructure shock occurs, the strategy stops trading the given ticker until it returns to its initial tick size. This is for simplicity, because the models would need a more complex fitting method to accommodate such an event. See Table 20 for the tick size rules.

Appendix B. Econometric Model Performance — Additional Results

Table 32. Alsayed and McGroarty [2] mid-quote direction predictions computed on six months of data for each ticker at Xetra.

	Accuracy (%)					**Potential Opportunities**				
Ticker \K^{AM}	$\boldsymbol{\delta}$	$\boldsymbol{2\delta}$	$\boldsymbol{3\delta}$	$\boldsymbol{4\delta}$	$\boldsymbol{5\delta}$	$\boldsymbol{\delta}$	$\boldsymbol{2\delta}$	$\boldsymbol{3\delta}$	$\boldsymbol{4\delta}$	$\boldsymbol{5\delta}$
ADS	71.7	70.1	66.1	64.5	64.2	207 093	51 518	22 702	13 452	9 151
ALV	74.8	64.1	70.6	75.6	76.1	41 165	1 120	316	172	138
BAS	75.1	75.2	66.6	64.4	63.4	227 108	28 800	9 027	5 218	3 428
BAYN	73.9	75.7	67.8	64.1	63.2	251 731	41 221	13 036	6 897	4 507
BEI	74.9	71.7	66.2	63.5	62.7	110 355	21 086	8 992	5 368	3 697
BMW	74.7	75.5	68.7	66.2	63.3	236 652	38 946	11 764	5 911	3 458
CBK	65.8	67.5	65.9	63.6	62.2	339 938	99 731	26 519	14 294	10 138
CON	70.6	70.0	67.2	65.3	64.1	222 551	71 282	31 413	17 958	11 853
DAI	73.5	72.8	67.1	64.1	62.3	440 231	91 651	34 176	18 742	11 843
DB1	69.1	67.8	66.0	63.5	61.1	201 577	62 592	24 100	13 997	9 530
DBK	73.1	73.5	69.0	65.0	63.6	416 043	61 684	15 043	6 782	3 895
DPW	73.6	68.0	65.7	66.7	70.1	41 428	4 792	2 201	1 026	435
DTE	66.6	69.9	70.0	69.1	69.0	430 101	68 517	9 369	4 015	2 495
EOAN	69.1	64.3	64.3	64.4	65.0	40 405	1 681	479	194	143
FME	73.1	68.4	64.7	63.0	62.3	114 853	20 998	10 183	6 552	4 590
FRE	71.3	71.2	68.1	64.8	63.6	197 753	64 364	30 590	17 911	12 534
HEI	72.8	71.0	66.9	65.5	64.7	174 606	38 452	14 460	8 213	5 385
HEN3	-	-	-	-	-	-	-	-	-	-
IFX	72.6	69.7	66.7	65.3	64.7	281 850	78 420	21 257	10 756	7 505
LHA	70.9	64.9	61.7	60.3	64.1	54 939	5 346	1 769	663	398
LIN	71.3	65.2	66.6	68.6	69.7	24 285	2 215	991	385	241
LXS	69.5	67.7	65.5	63.3	63.3	219 530	55 455	18 168	10 097	6 500
MRK	67.9	65.8	63.7	62.8	61.9	23 184	3 425	1 193	705	522
MUV2	73.2	63.2	58.6	60.7	60.2	33 239	2 010	691	239	166
RWE	72.2	69.8	64.8	61.6	58.8	171 798	23 412	7 991	4 267	2 295
SAP	73.6	71.5	67.1	66.0	63.3	145 500	19 100	7 140	3 957	2 164
SIE	75.1	74.5	66.1	63.6	63.0	243 347	32 407	10 405	6 063	3 999
TKA	68.2	64.9	61.2	59.0	63.3	73 814	7 277	2 045	791	387
VOW3	74.1	67.1	60.4	60.5	64.6	77 109	6 021	1 691	580	285

Table 33. Alsayed and McGroarty [2] mid-quote direction predictions computed on six months of data for each ticker at BATS.

	Accuracy (%)					**Potential Opportunities**				
Ticker \K^{AM}	$\boldsymbol{\delta}$	$\boldsymbol{2\delta}$	$\boldsymbol{3\delta}$	$\boldsymbol{4\delta}$	$\boldsymbol{5\delta}$	$\boldsymbol{\delta}$	$\boldsymbol{2\delta}$	$\boldsymbol{3\delta}$	$\boldsymbol{4\delta}$	$\boldsymbol{5\delta}$
ADS	72.5	72.4	69.2	67.9	67.8	186 741	44 983	18 573	10 650	7 115
ALV	65.2	70.7	78.1	81.2	80.0	40 214	1 040	319	191	160
BAS	76.6	77.1	70.2	68.7	68.4	226 470	26 610	7 384	4 171	2 741
BAYN	75.8	76.9	70.4	67.4	67.3	244 096	37 637	10 625	5 164	3 258
BEI	73.2	75.0	71.8	69.5	69.5	108 102	19 604	7 770	4 366	2 817
BMW	72.5	72.9	68.7	67.9	65.2	211 945	34 597	9 453	4 399	2 439
CBK	60.5	64.1	64.8	64.8	64.0	314 401	88 256	23 712	12 619	8 984
CON	71.4	71.1	69.3	68.0	67.0	159 198	52 819	22 658	12 242	7 924
DAI	70.5	70.3	66.2	64.2	63.6	429 938	86 667	30 589	16 030	9 785
DB1	69.4	70.4	70.6	69.6	68.6	161 937	49 272	18 013	9 583	6 132
DBK	76.2	76.9	72.6	69.5	67.8	410 400	61 851	14 984	6 539	3 671
DPW	71.8	70.4	69.6	70.4	69.9	41 742	4 160	1 812	866	415
DTE	64.4	68.6	68.3	65.9	64.9	484 711	71 700	9 915	4 351	2 757
EOAN	70.1	72.5	72.8	72.2	71.4	39 904	1 567	463	219	168
FME	70.9	70.8	68.5	67.6	68.1	119 564	18 461	8 152	4 981	3 460
FRE	69.3	69.8	68.8	67.7	67.3	128 239	43 937	19 725	10 732	7 229
HEI	68.3	70.4	70.4	69.1	67.9	132 583	31 014	10 982	6 029	4 005
HEN3	69.1	68.5	67.2	65.9	66.0	116 053	27 137	10 051	5 159	3 185
IFX	66.4	67.4	66.4	65.2	64.8	248 676	72 051	19 118	9 509	6 624
LHA	74.0	71.4	68.6	65.7	68.6	54 420	4 793	1 323	499	290
LIN	70.0	70.9	70.2	68.8	67.4	24 303	1 860	805	336	239
LXS	68.3	69.8	70.3	70.3	70.3	164 776	42 692	13 558	7 199	4 644
MRK	68.2	69.5	67.0	68.1	67.7	24 930	3 094	932	508	365
MUV2	73.0	75.4	70.3	65.5	64.4	33 414	1 775	583	226	160
RWE	-	-	-	-	-	-	-	-	-	-
SAP	74.0	74.1	71.4	69.8	67.2	145 724	17 172	5 822	3 059	1 688
SIE	71.8	75.5	71.0	69.3	68.7	258 849	30 501	8 470	4 527	2 880
TKA	72.1	71.8	71.5	67.5	65.1	75 764	6 849	1 714	667	350
VOW3	71.7	72.7	69.3	68.8	60.7	80 921	5 482	1 361	468	257

Table 34. ADLMLR bid price process direction predictions computed on six months of data for each ticker at Xetra and multiple Ks.

	Accuracy (%)								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	81.8	81.6	81.0	80.3	79.4	78.7	77.5	74.4	71.5
ALV	77.7	77.7	74.1	71.6	71.9	70.1	68.6	68.0	67.1
BAS	86.1	86.0	85.6	84.8	83.8	82.6	81.4	80.2	79.1
BAYN	87.2	86.8	86.0	85.2	84.3	83.4	82.4	81.4	80.5
BEI	88.1	87.8	87.5	87.0	86.6	85.8	85.2	84.4	83.6
BMW	88.2	87.9	87.6	86.9	86.1	85.1	84.1	83.0	81.9
CBK	82.5	82.1	81.3	79.8	77.0	69.2	69.7	71.3	72.4
CON	79.8	79.8	79.7	79.5	78.9	78.3	77.5	77.3	76.1
DAI	84.5	84.5	84.3	83.8	83.2	82.5	81.6	80.7	79.7
DB1	80.7	80.5	80.1	79.9	79.2	78.4	77.4	76.6	71.8
DBK	88.6	86.4	86.1	85.8	85.2	84.6	84.0	83.3	82.4
DPW	86.4	86.6	85.8	85.2	83.9	82.6	81.0	79.9	78.9
DTE	88.2	88.1	87.7	87.1	86.3	85.4	84.4	83.4	82.5
EOAN	85.9	84.7	82.7	80.3	76.5	74.4	72.8	71.3	70.0
FME	84.9	85.2	85.0	84.7	84.2	83.7	83.0	82.3	81.5
FRE	80.5	80.4	79.4	76.7	68.6	68.9	69.2	70.1	70.8
HEI	85.5	85.4	85.0	84.7	84.1	83.5	82.6	81.7	80.9
HEN3	-	-	-	-	-	-	-	-	-
IFX	85.9	85.9	85.7	85.5	85.2	84.9	84.7	84.2	83.7
LHA	85.3	85.5	85.0	84.6	84.0	83.2	82.4	81.3	80.4
LIN	81.4	81.6	80.9	79.3	76.5	75.0	73.4	71.8	70.7
LXS	81.1	81.5	81.8	81.5	81.3	80.6	80.1	79.5	78.7
MRK	82.4	82.8	82.3	81.5	80.8	80.3	79.8	78.8	78.5
MUV2	82.3	82.3	81.1	79.6	77.7	75.4	73.0	71.2	69.9
RWE	85.2	85.2	84.8	84.3	83.5	82.6	81.7	80.8	79.6
SAP	84.8	84.6	84.0	83.2	82.2	81.1	79.9	78.6	77.4
SIE	87.3	87.4	86.7	86.1	85.1	84.0	82.8	81.9	80.8
TKA	86.5	86.7	86.3	85.7	84.9	84.0	83.0	81.6	80.4
VOW3	86	85.0	84.0	82.6	80.7	79.5	78.1	76.7	75.8
	Potential Opportunities								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	30 921	35 286	39 968	45 093	50 575	50 272	28 165	10 629	4 297
ALV	283	394	575	980	1 663	2 743	4 032	5 601	7 281
BAS	18 160	23 046	28 044	33 558	39 700	46 766	54 954	64 071	73 405
BAYN	26 195	31 442	36 851	42 762	49 519	57 000	65 163	73 748	82 099
BEI	15 945	18 433	21 030	23 664	26 402	29 328	32 213	35 120	37 811
BMW	25 793	30 810	35 998	41 453	47 254	53 655	60 483	67 707	74 969
CBK	33 676	40 120	41 088	35 718	31 174	10 751	12 381	11 114	12 799
CON	26 077	29 178	32 502	36 021	39 907	43 920	48 190	38 752	23 615
DAI	54 762	63 187	71 882	80 629	90 099	99 954	110 593	121 646	132 795
DB1	19 627	22 824	26 454	30 525	35 270	40 394	45 854	35 404	7 226
DBK	36 681	34 372	42 140	50 664	59 908	70 333	81 715	94 324	107 699
DPW	3 406	4 140	5 000	6 233	7 871	9 861	12 165	14 388	16 578
DTE	15 898	21 367	27 264	33 409	39 635	46 320	53 142	60 140	67 593
EOAN	1 266	1 755	2 364	3 463	5 560	8 698	13 005	17 242	20 964
FME	12 382	14 521	16 831	19 370	22 244	25 238	28 402	31 444	34 313
FRE	28 444	26 874	24 332	16 284	5 360	6 136	6 845	4 108	4 861
HEI	18 116	20 956	23 910	26 966	30 117	33 293	36 637	40 151	43 507
HEN3	-	-	-	-	-	-	-	-	-
IFX	31 499	36 515	41 742	47 484	54 301	62 839	72 185	80 774	83 824
LHA	6 115	7 515	8 892	10 548	12 648	15 196	17 983	20 867	23 714
LIN	1 422	1 767	2 166	2 739	3 705	5 053	6 810	8 821	10 925
LXS	11 547	13 920	16 593	19 606	23 095	26 695	30 538	34 767	39 377
MRK	1 657	2 028	2 490	3 040	3 691	4 414	5 135	5 893	6 539
MUV2	1 145	1 493	1 866	2 428	3 318	4 573	6 239	8 210	10 290
RWE	17 519	21 529	25 494	29 483	34 003	38 907	44 709	51 344	58 359
SAP	13 213	15 657	18 212	21 188	24 919	29 452	34 764	40 652	46 600
SIE	18 183	22 674	27 466	32 574	38 546	45 283	52 742	60 341	67 885
TKA	5 462	7 323	9 218	11 086	12 822	14 743	16 816	19 029	21 456
VOW3	5 593	6 848	8 296	10 197	12 599	15 432	18 646	22 050	25 340

Table 35. ADLMLR bid price process direction predictions computed on six months of data for each ticker at BATS and multiple Ks.

	Accuracy (%)								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	76.0	76.3	76.5	76.1	75.6	75.0	74.1	73.9	73.4
ALV	79.3	79.2	79.0	78.9	78.4	78.0	77.5	76.7	75.8
BAS	79.3	79.7	79.6	79.3	78.9	78.1	77.2	76.2	75.1
BAYN	78.9	79.1	79.0	78.7	78.2	77.7	77.1	76.2	75.2
BEI	80.1	79.7	79.1	78.3	77.3	76.1	74.3	72.3	69.6
BMW	79.2	79.1	79.0	78.9	78.5	78.3	77.9	77.3	76.5
CBK	69.4	70.5	71.5	71.9	72.1	71.9	71.8	71.4	71.2
CON	75.4	75.4	75.5	75.4	75.2	74.9	75.4	74.4	74.1
DAI	74.6	74.6	74.6	74.3	74.0	73.6	73.1	72.3	72.2
DB1	78.2	78.0	78.0	77.8	77.3	76.9	76.5	75.6	74.8
DBK	86.2	82.9	83.3	83.3	83.3	83.1	82.7	82.3	81.8
DPW	77.0	76.8	76.5	75.7	74.6	73.3	71.9	70.6	69.1
DTE	83.9	83.4	83.0	82.3	81.6	80.6	79.7	78.6	77.5
EOAN	84.2	83.1	82.3	81.4	80.4	79.4	79.2	78.4	77.7
FME	77.0	76.9	76.8	76.3	75.6	74.3	72.9	71.4	70.7
FRE	78.3	78.0	77.5	76.9	73.9	75.9	72.0	71.9	72.2
HEI	79.8	79.3	78.9	78.0	77.3	76.6	75.8	74.8	73.7
HEN3	76.7	76.6	76.1	75.3	74.5	73.9	73.1	72.1	72.9
IFX	78.3	78.2	78.0	77.6	77.0	76.1	74.8	72.2	69.5
LHA	86.7	87.1	87.1	86.2	85.5	84.8	84.1	83.2	82.4
LIN	78.7	78.3	78.5	78.3	78.0	77.5	76.9	76.2	75.2
LXS	80.5	80.9	80.8	81.0	81.0	80.7	79.9	78.9	78.1
MRK	75.8	75.8	75.2	74.7	73.3	72.0	70.7	69.7	68.3
MUV2	79.8	79.2	79.3	78.8	78.1	77.6	76.9	76.1	74.9
RWE	-	-	-	-	-	-	-	-	-
SAP	78.4	78.1	77.7	77.2	76.5	75.9	75.0	73.6	71.7
SIE	78.0	78.1	77.8	77.3	76.7	75.8	74.6	73.3	72.1
TKA	84.8	84.9	84.1	83.9	83.3	82.6	81.6	80.6	79.4
VOW3	76.9	76.2	75.4	74.3	73.4	72.4	71.6	70.9	70.0
	Potential Opportunities								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	16 658	19 906	23 755	28 191	30 847	33 653	39 495	23 346	7 419
ALV	4 870	6 344	7 446	8 206	8 709	9 071	9 351	9 677	10 098
BAS	20 357	24 871	30 271	36 608	43 650	51 218	59 218	68 748	79 569
BAYN	25 715	31 102	37 115	43 786	51 053	59 118	68 008	77 702	87 659
BEI	14 178	16 890	19 827	23 281	27 085	26 959	29 558	31 972	31 720
BMW	16 988	20 273	24 103	28 490	33 255	38 726	44 826	51 328	58 495
CBK	6 026	7 875	9 899	12 208	14 584	17 098	19 777	22 585	25 456
CON	18 493	21 154	24 325	27 894	31 800	36 206	7 939	7 339	7 502
DAI	27 757	32 729	38 415	44 663	51 605	59 466	68 222	78 085	62 966
DB1	7 293	8 680	10 281	12 158	14 211	16 655	19 480	22 834	26 717
DBK	30 084	26 839	33 668	41 185	49 402	58 302	68 056	78 582	89 943
DPW	6 952	8 389	9 972	11 647	13 573	15 775	17 121	18 375	20 361
DTE	19 527	24 681	30 420	36 474	42 941	49 656	56 636	63 938	71 807
EOAN	2 091	4 359	7 070	9 811	12 817	14 278	15 315	16 757	18 021
FME	10 772	13 039	15 681	18 746	22 141	26 056	30 596	35 639	33 514
FRE	16 801	19 567	21 713	25 015	18 485	8 513	4 575	2 574	3 163
HEI	10 130	11 770	13 553	15 539	17 796	20 102	22 509	25 152	28 017
HEN3	12 073	13 663	15 491	17 561	19 849	22 378	20 441	21 923	20 310
IFX	24 511	29 214	34 819	41 175	48 318	56 331	58 735	56 609	37 044
LHA	3 906	5 705	7 584	9 524	11 398	13 261	15 265	17 372	19 775
LIN	5 484	7 013	8 264	9 260	10 026	10 618	11 162	11 695	12 327
LXS	4 135	5 163	6 347	7 658	9 341	11 274	13 570	16 139	19 026
MRK	3 030	3 581	4 132	4 626	5 181	5 722	6 369	6 976	7 707
MUV2	3 617	5 042	6 265	7 312	8 176	8 885	9 537	10 194	10 932
RWE	-	-	-	-	-	-	-	-	-
SAP	15 259	19 120	22 756	26 421	30 150	34 157	38 404	43 569	50 664
SIE	16 256	20 633	25 579	31 271	37 625	44 290	51 478	58 937	66 880
TKA	4 009	5 651	7 628	9 474	11 436	13 528	15 714	17 975	20 742
VOW3	9 527	12 781	16 099	19 535	22 849	25 774	28 312	30 527	32 671

Table 36. ADLMLR ask price process direction predictions computed on six months of data for each ticker at Xetra and multiple Ks.

	Accuracy (%)								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	80.9	80.7	80.3	79.8	79.1	78.1	77.2	76.3	71.6
ALV	76.7	75.9	75.1	73.2	70.2	68.9	67.5	66.3	65.6
BAS	86.4	86.2	85.6	84.7	83.4	82.2	80.9	79.7	78.5
BAYN	84.8	84.5	84.0	83.5	82.7	82.0	81.0	79.9	78.8
BEI	86.8	86.7	86.3	85.9	85.2	84.6	83.8	82.9	82.4
BMW	87.9	87.7	87.3	86.7	85.9	84.8	83.8	82.7	81.6
CBK	82.3	72.1	72.2	72.3	72.7	72.5	72.4	72.3	72.3
CON	80.1	80.1	80.1	79.9	79.5	78.8	78.0	76.7	73.9
DAI	84.6	84.5	84.3	83.8	83.3	82.5	81.7	80.7	79.8
DB1	80.5	80.3	79.8	79.5	78.9	78.4	77.6	76.8	71.9
DBK	86.1	86.1	85.7	85.4	84.8	84.2	83.4	82.6	81.8
DPW	85.2	85.0	84.5	83.8	82.3	80.9	79.7	78.7	77.9
DTE	89.7	89.5	89.0	88.4	87.6	86.4	85.2	84.2	83.2
EOAN	85.8	85.3	83.2	80.5	77.0	74.6	72.3	70.5	69.0
FME	83.8	84.2	84.2	84.2	83.8	83.2	82.8	82.3	81.7
FRE	77.2	77.4	76.4	75.1	72.0	71.2	68.7	69.5	69.9
HEI	85.6	85.6	85.4	85.1	84.6	83.9	83.0	82.2	81.2
HEN3	-	-	-	-	-	-	-	-	-
IFX	86.4	86.2	85.9	85.6	85.2	84.9	84.6	84.3	83.8
LHA	86.5	86.3	85.8	85.4	84.6	83.6	82.7	81.8	80.8
LIN	81.4	80.7	79.2	77.4	75.5	73.9	72.4	71.0	69.8
LXS	81.0	81.4	81.4	81.4	81.1	80.6	80.2	79.6	78.8
MRK	78.5	78.3	78.9	79.6	78.8	78.4	77.7	77.5	76.9
MUV2	82.6	82.2	81.9	80.5	79.3	76.5	74.5	72.9	70.8
RWE	83.4	83.4	83.4	82.9	82.1	81.2	80.3	79.3	78.4
SAP	85.2	84.9	84.1	83.1	81.8	80.7	79.5	78.3	77.2
SIE	86.6	86.8	86.5	85.9	84.9	83.9	83.0	82.0	81.1
TKA	85.8	85.9	85.6	85.0	84.1	83.2	82.1	81.1	79.8
VOW3	85.6	85.7	85.2	83.9	82.5	81.0	79.4	77.9	76.8
	Potential Opportunities								
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	28 608	32 698	37 266	42 006	47 189	52 778	58 452	29 678	7 789
ALV	643	937	1 511	2 590	4 295	6 461	9 008	11 716	14 293
BAS	20 883	25 677	30 490	35 730	42 148	49 724	58 185	67 347	76 681
BAYN	21 160	25 974	31 020	36 515	42 747	49 702	57 503	66 176	75 580
BEI	17 641	20 042	22 644	25 499	28 429	31 471	34 480	37 391	35 580
BMW	24 655	29 771	34 987	40 413	46 077	52 271	58 846	65 856	73 228
CBK	13 372	4 758	5 655	6 614	7 666	8 691	9 886	11 214	12 760
CON	25 429	28 756	32 516	36 670	41 169	45 994	40 291	21 207	3 126
DAI	55 247	64 215	73 575	83 238	93 416	104 200	116 037	128 391	126 253
DB1	19 476	22 574	26 268	30 435	35 137	40 248	45 788	32 260	9 796
DBK	27 432	34 604	42 219	50 449	59 533	69 492	80 640	92 856	105 613
DPW	2 934	3 602	4 314	5 311	6 789	8 567	10 623	12 804	14 900
DTE	19 116	24 869	30 456	36 400	42 540	49 112	56 183	63 660	71 108
EOAN	1 283	1 848	2 560	3 549	5 301	7 866	10 960	14 473	18 096
FME	11 992	14 147	16 433	19 057	22 004	25 328	28 525	31 738	34 802
FRE	20 999	22 862	21 715	13 417	5 556	6 314	2 361	2 959	3 495
HEI	18 568	21 667	24 912	28 208	31 675	35 262	38 996	42 720	46 319
HEN3	-	-	-	-	-	-	-	-	-
IFX	31 910	36 935	41 767	47 211	53 824	62 397	71 723	80 246	86 962
LHA	6 181	7 661	9 079	10 863	13 014	15 616	18 424	21 423	24 285
LIN	1 369	1 675	2 108	2 804	3 827	5 208	6 891	8 716	10 637
LXS	10 966	13 216	15 819	18 753	22 050	25 553	29 378	33 623	38 364
MRK	2 634	3 102	3 141	3 549	4 245	5 020	5 765	6 468	7 097
MUV2	1 039	1 321	1 658	2 109	2 714	3 636	4 978	6 625	8 563
RWE	16 029	19 631	23 356	27 363	31 711	36 679	42 362	48 561	55 273
SAP	13 183	15 591	18 104	21 135	25 025	29 603	34 963	40 685	46 425
SIE	18 078	22 682	27 290	32 468	38 319	44 929	52 046	59 466	66 941
TKA	6 469	8 270	10 084	11 851	13 679	15 658	17 771	20 199	22 816
VOW3	3 904	5 013	6 229	7 706	9 501	11 683	14 275	17 308	20 482

Table 37. ADLMLR ask price process direction predictions computed on six months of data for each ticker at BATS and multiple Ks.

					Accuracy (%)				
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	75.5	75.6	75.4	75.0	74.3	73.6	73.5	72.9	71.1
ALV	78.1	78.5	78.1	77.8	77.2	76.9	76.4	75.9	75.1
BAS	79.8	79.9	79.5	79.1	78.3	77.4	76.5	75.2	74.0
BAYN	78.9	78.8	78.7	78.4	78.0	77.4	76.7	75.9	75.0
BEI	79.3	79.0	78.5	78.0	77.3	76.6	75.5	74.0	71.8
BMW	79.2	79.1	79.1	78.7	78.3	77.6	76.8	76.5	76.8
CBK	69.3	70.2	71.3	71.2	71.5	71.6	71.4	71.3	71.2
CON	75.3	75.5	75.5	75.6	75.7	76.6	76.7	76.8	76.4
DAI	75.2	75.2	75.2	75.1	74.9	74.6	74.1	73.3	72.8
DB1	79.0	79.1	78.5	77.9	77.1	76.4	75.6	74.4	73.2
DBK	82.5	82.5	82.5	82.5	83.5	83.1	82.8	82.4	81.8
DPW	77.6	77.3	77.0	76.1	74.8	73.4	71.8	70.3	68.6
DTE	82.4	82.4	82.0	81.5	80.8	79.9	78.7	80.2	78.9
EOAN	80.9	81.1	80.6	80.1	79.2	78.2	77.2	76.1	74.7
FME	75.8	76.1	76.3	76.2	75.6	74.9	73.8	72.3	70.9
FRE	77.2	77.2	77.1	76.6	74.9	74.5	74.7	73.9	74.0
HEI	79.3	78.9	78.4	77.8	77.1	76.3	75.3	74.3	73.3
HEN3	77.2	76.7	76.2	75.6	74.5	73.6	72.4	71.6	70.6
IFX	79.2	79.1	78.6	78.1	77.5	76.3	74.8	73.0	70.3
LHA	87.4	87.7	87.3	86.9	86.3	85.7	84.6	83.7	82.7
LIN	77.5	78.2	78.4	78.2	77.8	77.4	77.0	76.4	75.6
LXS	80.5	80.3	79.6	79.0	78.2	77.2	75.7	74.3	73.2
MRK	76.8	76.4	75.4	74.8	74.5	73.5	72.1	70.6	68.9
MUV2	80.3	81.4	80.5	80.0	79.2	78.8	78.4	77.8	77.1
RWE	-	-	-	-	-	-	-	-	-
SAP	78.8	78.6	78.2	77.7	76.9	76.1	75.3	74.0	72.7
SIE	77.5	77.4	77.2	76.7	75.9	74.9	73.8	72.6	71.3
TKA	84.7	84.8	84.1	83.5	82.8	81.9	80.9	79.7	78.6
VOW3	76.3	75.6	75.4	74.6	73.7	72.8	71.8	70.9	69.8
					Potential Opportunities				
Ticker \K	**Peak**	**Peak - 0.025**	**Peak - 0.050**	**Peak - 0.075**	**Peak - 0.100**	**Peak - 0.125**	**Peak - 0.150**	**Peak - 0.175**	**Peak - 0.200**
ADS	20 460	24 126	28 180	32 813	38 273	44 420	24 071	10 687	7 498
ALV	3 747	5 078	6 145	6 867	7 452	7 811	8 105	8 411	8 787
BAS	25 399	30 853	37 300	44 306	51 788	59 716	68 571	78 824	90 271
BAYN	26 801	32 133	38 087	44 553	51 822	59 762	68 834	78 496	88 362
BEI	12 084	14 542	17 175	20 086	23 480	27 283	31 779	33 811	33 689
BMW	21 650	25 444	29 913	34 879	40 243	46 361	52 962	55 564	50 812
CBK	4 992	5 985	7 238	8 611	10 063	11 543	13 101	14 694	16 337
CON	15 444	17 739	20 425	23 540	26 320	23 886	20 533	9 466	9 026
DAI	32 913	38 482	45 002	52 334	60 786	70 333	81 285	93 833	74 003
DB1	10 432	12 211	14 248	16 478	19 200	22 245	25 835	29 944	32 847
DBK	21 476	28 297	35 242	42 448	37 368	44 493	51 933	60 041	68 775
DPW	6 143	7 524	8 941	10 453	12 158	14 026	16 268	17 189	16 961
DTE	14 109	18 768	24 218	30 079	36 409	43 447	52 738	37 771	43 339
EOAN	2 993	5 818	8 898	11 518	13 743	15 553	17 149	18 670	20 339
FME	9 976	12 049	14 631	17 657	21 001	24 838	29 117	33 994	36 397
FRE	17 959	20 639	23 794	27 208	24 996	5 228	6 310	3 903	4 625
HEI	9 777	11 465	13 325	15 292	17 544	19 940	22 604	25 513	28 574
HEN3	13 097	14 941	17 055	19 271	18 646	21 219	21 552	22 979	13 939
IFX	25 927	30 711	36 085	42 126	48 862	54 363	57 507	59 465	39 525
LHA	4 736	6 846	9 041	11 155	13 133	15 062	17 074	19 219	21 531
LIN	4 855	6 385	7 677	8 748	9 590	10 188	10 670	11 131	11 693
LXS	8 759	10 471	12 631	15 159	18 011	21 241	25 009	29 290	21 101
MRK	3 158	3 772	4 388	4 905	5 328	5 817	6 357	6 991	7 726
MUV2	2 001	2 917	3 858	4 655	5 235	5 663	5 971	6 209	6 490
RWE	-	-	-	-	-	-	-	-	-
SAP	19 983	24 210	28 468	32 327	35 884	39 382	42 942	46 889	51 624
SIE	17 497	21 767	26 841	32 650	39 001	46 036	53 309	60 974	69 181
TKA	5 608	7 616	9 624	11 699	13 721	15 863	18 062	20 423	22 864
VOW3	7 942	10 377	13 140	16 190	19 353	22 387	25 298	27 822	30 296

Appendix C. Execution Rules

1. Each limit order has a standing quantity that must be executed before the order is executed.
2. That standing quantity is computed from the following steps:
    a. If the limit order's price of a buy/sell order is equal to the best bid/ask price, the order's standing quantity becomes the current best bid/ask volume.
    b. If the limit price of a buy/sell order is below/above the best bid/ask price, the order's standing quantity is undefined. In that instance, the trading and quoting emulator waits for the limit order's price to be equal to the best bid/ask price and it sets the standing quantity according to 2.a.
    c. If the limit order's buy/sell price is above/below the best bid/ask price, the order is filled.
    d. If the standing quantity has been defined for a limit order, it can only be changed by a future execution.
3. A limit order can be executed by a trade occurring at the limit order's price. The standing quantity must be executed first. If it has been executed completely, then the limit order can be executed. If the remaining trade size is not large enough to fill the limit order's size, then a partial filling occurs. Limit orders with an undefined standing quantity cannot be executed by a trade.
4. A limit order can be executed when the best ask/bid price becomes lower/greater than the buy/sell limit order's price. This also holds for limit orders with undefined standing quantities.
5. A limit order is filled when the best bid/ask price becomes lower/greater than the buy/sell limit order's price, regardless of its standing quantity. This also holds for limit orders with undefined standing quantities.

The trading and quoting emulator is conservative in some regards, especially considering the static standing quantity that must be executed before the corresponding limit order, because it ignores cancellations decreasing that quantity after the order has been placed, which follows from rules 1 and 2.a. Also, whenever a limit order is placed deeper than LOB level 1 and its price becomes the top of the book after some time, the limit order is put at the end of the queue of all the orders also at the new level 1 regardless of its actual position in that queue, which follows from rules 2.a and 2.b.

Third Article.

Deep Unsupervised Anomaly Detection in High-Frequency Markets

by

Cédric Poutré[1], Manuel Morales[1], and Didier Chételat[2]

([1]) Department of Mathematics and Statistics
Université de Montréal

([2]) Canada Excellence Research Chair in Data Science for Real Time Decision Making
Polytechnique Montréal

This article was submitted to The Journal of Finance and Data Science.

The main contributions of Cédric Poutré for this article are presented.

- Creation of the entire model;
- Design of the trade–based fraud simulator;
- Production of all numerical results;
- Writing most of the manuscript.

Both Manuel Morales and Didier Chételat helped with the manuscript.

Abstract. Inspired by recent advances in the deep learning literature, this article introduces a novel hybrid anomaly detection framework specifically designed for limit order book (LOB) data. A modified Transformer autoencoder architecture is proposed to learn rich temporal LOB subsequence representations, which eases the separability of normal and fraudulent time series. A dissimilarity function is then learned in the representation space to characterize normal LOB behavior, enabling the detection of any anomalous subsequences out-of-sample. We also develop a complete trade–based manipulation simulation methodology able to generate a variety of scenarios derived from actual trade–based fraud cases. The complete framework is tested on LOB data of five NASDAQ stocks in which we randomly insert synthetic quote stuffing, layering, and pump-and-dump manipulations. We show that the proposed asset–independent approach achieves new state-of-the-art fraud detection performance, without requiring any prior knowledge of manipulation patterns.

Keywords: Limit order book; Time series anomaly detection; Deep learning; Trade–based manipulation; Dissimilarity model; Unsupervised learning

1. Introduction

Exchange regulators, who constantly monitor markets to unveil potential manipulations, traditionally perform their investigation manually. When a potentially fraudulent event, or sequence of events, is detected by the automated system, market analysts have the responsibility to carry on the necessary research and analysis to conclude whether or not there has been a violation to the exchange's rules and regulations. The enormous volume of orders means that this task is especially laborious, and investigations can take years.[28] A first solution to this problem is to implement rule–based systems that can automatically flag orders as suspicious. In fact, current market regulators' systems are based on deterministic rules inferred from a set of known delinquent patterns defined by experts (Golmohammadi et al. [21]), on which we shall return. However, such systems have seen limited success in practice, as it is difficult to completely formalize abnormal behaviors by a rule–based system, because defining all anomalies in a trading context is not realistically feasible.

[28] SEC's Division of Enforcement (accessed March 18, 2023).

A more promising avenue for this kind of problem is based in machine learning techniques, which have seen a lot of success in a variety of real–world applications (Pang et al. [40]). Moreover, the static nature of rule–based systems is fundamentally ill-matched with the dynamic nature of financial markets, where fraudulent patterns might be constantly evolving with the market (Lin [33]). In contrast, machine learning approaches can dynamically learn unusual order patterns over time, adapting to evolving market conditions. But the current literature on machine learning in financial market manipulation lacks generality, as only very limited sets of features and/or fraud types are studied simultaneously. A more generic framework able to detect several types of fraudulent patterns by utilizing a larger set of information would be valuable (Khodabandehlou and Golpayegani [26]). We aim to fill this gap by proposing a new model based on recent state-of-the-art methods in the deep learning literature capable of managing multivariate time series. Furthermore, previous papers rely on repeating the same limited sets of fraudulent orders to evaluate their methods, probably overestimating their capabilities. We also address that problem with a more exhaustive simulation approach generating further representative sets of trade–based manipulations. Hence, the detection results presented in this paper are more faithful to what could be achieved in practice..

The ultimate objective of any financial market fraud detection system, whether human, rule or machine learning–based, is the detection of all *trade–based manipulations*. Trade–based manipulations are defined as "[...] a type of behavior [that] consists of effecting transactions or orders to trade which (a) give, or are likely to give a false or misleading impression as to the supply of, or demand for, or as to the price of one or more [qualifying investments] or (b) secure the price of one or more such investments at an abnormal or artificial level."[29] Multiple manipulation schemes fit that description: advancing the bid, reducing the ask, wash sales, marking the close, pump-and-dump, layering/spoofing, quote stuffing, and so on, all with their respective footprint.[30] In this context, anomalies are sequences of market

[29] Financial Conduct Authority, MAR 1.6.1 Market abuse (manipulation transactions) FCA Handbook (accessed March 18, 2023).

[30] See Siering et al. [53] for a complete taxonomy of financial market manipulations.

events that are out of the ordinary, and that could potentially be associated with trade–based manipulations. Market events include new order submissions, cancellations or modifications of a past order, and executions. All of which are recorded sequentially in central limit order books (LOBs), forming a collection of outstanding limit orders. The LOB depicts the prices at which market participants are willing to buy and sell a given asset, as well as the volume available at each price point, at any given time during trading hours. We note that, although frauds are anomalous in well regulated markets, not all anomalies are necessarily frauds. For example, unusual external events might trigger perfectly legal, but unusual, behavior on markets.

The ideal situation would be to develop a supervised machine learning model to classify orders as part of a manipulation tactic. Unfortunately, precisely because investigations often take years, and fraud methods evolve quickly, most markets have too few examples of such trading activities to reasonably think about using supervised machine learning methods. A more realistic alternative and useful approach would be to develop unsupervised anomaly detection techniques for financial markets. Such methods could flag specific subsequences of orders as suspicious, which could then be further investigated by market regulators, reducing the burden of their analysis. This challenge has two objectives. First and foremost, finding a suitable unsupervised anomaly detection method that can flag most, if not all, true cases of trade–based manipulation as anomalous, i.e., high recall. Second, the method should report as few non-fraudulent anomalies as possible, i.e., high precision, to limit the costs associated with the analysis non-fraudulent orders.

In this paper, we propose a novel approach specifically tailored to high-frequency financial markets that performs better than competing methods. The approach is twofold. First, an unsupervised autoencoder based on the Transformer architecture of Vaswani et al. [57] is trained on LOB–based features specifically built to capture a multitude of trade–based manipulations. We empirically show that the autoencoder learns temporal representation vectors useful in characterizing LOB subsequences, thus easing the separability of fraudulent patterns from regular subsequences. Second, a discriminative model estimates the boundary between normal and abnormal autoencoder representations, creating a dissimilarity function that enables the detection of fraudulent orders out-of-sample. This hybrid approach reduces the gap between state-of-the-art deep learning methods, and financial market fraud detection research. We show that our method, which does not make use of any fraud examples, tends to both capture all true frauds, and has lower false positive rate than competing unsupervised methods on the LOBSTER data set (Huang and Polak [23]). The ability of the framework to detect diverse types of manipulations is an important step in the advancement of stock market fraud detection literature (Khodabandehlou and Golpayegani [26]). Hence, we also propose a more exhaustive trade–based manipulation simulation methodology able to generate multiple fraud scenarios, setting us further apart from previous literature, and rendering our results more reliable. Finally, for the first time in the literature, we also quantitatively study the complexity of detecting certain types of trade–based manipulations, thus providing a new comparative standard other than the conventional anomaly detection metrics.

The paper is divided as follows. Section 2 introduces the literature on machine learning and deep learning anomaly detection models, with a focus on time series methods. It also provides a review of trade–based manipulation detection techniques put forward in the financial literature. Section 3 presents the unsupervised hybrid deep learning framework proposed to capture anomalous behavior in LOB time series. Section 4 details the financial data used, and describes three popular trade–based manipulation techniques that are then simulated to quantify the framework's performance. The section ends by presenting the proposed LOB features to capture fraudulent patterns. Section 5 carries out the numerical experiments and analyzes the effectiveness of the methodology on simulated frauds, and Section 6 concludes the paper.

2. Literature review

Anomaly detection is an active subfield of machine learning successful in a plethora of industrial applications (see Agrawal and Agrawal [3] for a comprehensive overview). Following the nomenclature of Blázquez-García et al. [5] and Chalapathy and Chawla [8], the problem addressed in this paper can be classified as multivariate collective outlier detection, a niche outlier type which limits the pool of applicable detection techniques. Indeed, we are interested in identifying sets of orders that jointly behave unusually, but are otherwise normal on an individual basis. Blázquez-García et al. [5] describe two types of subsequence outlier detection methods proposed in the literature: model–based, and dissimilarity–based. The first family of techniques tries to find subsequences that strongly deviate from a model's expected value, either by prediction or by estimation, whereas the second family aims to detect subsequences' representations that stray from a reference of normalcy. Predicting the stock market is a notoriously challenging task (Gandhmal and Kumar [19]), meaning that prediction–based models would be hazardous for anomaly detection on LOB data. This leaves only estimation and dissimilarity methods as suitable anomaly detectors in this context. The proposed framework falls in the second family, for which the literature is notably scarce.

Representing time series of LOBs necessitates a great ability in capturing temporal and spatial information, because of its complexity. Deep learning models have made great strides in that sense, especially for large data sets, which explains the growing interest in deep anomaly detection methods. Recent techniques include autoencoders (AEs): MSCRED (Zhang et al. [62]), OmniAnomaly (Su et al. [55]), USAD (Audibert et al. [4])); Generative adversarial networks (GANs): MAD-GAN (Li et al. [32]); Graph neural networks (GNNs): MTDAD-GAT (Zhao et al. [63]), GDN (Deng and Hooi [15]); Deep one-class networks: THOC (Shen et al. [52]). Most papers follow an unsupervised approach based on recurrent networks, e.g., long short-term memory networks (LSTMs, Hochreiter and Schmidhuber [22]). The unsupervised approach hinges on the assumption that anomalies in the data are either rare or absent, and that models are learning usual system behaviors well enough to distinguish outliers out-of-sample. The Transformer architecture of Vaswani et al. [57] has also started to appear in the time series anomaly detection literature (Meng et al. [35], Xu et al. [61]), motivated by its success in natural language processing tasks, and its greater memory and parallelization capabilities compared to recurrent models. Finally, hybrid approaches combining self-supervised representation learning and one-class classifiers, such as kernel density estimation (KDE, Parzen [41]), one-class support vector machine (OC-SVM, Schölkopf et al. [51]) or k-nearest neighbors (kNN, Cover and Hart [14]) recently achieved state-of-the-art unsupervised anomaly detection performance on the visual domain (Reiss et al. [43, 44]; Sohn et al. [54]). Inspired by these recent advances, we propose an unsupervised hydrid methodology combining the temporal context representation capabilities of Transformers, and the effectiveness of statistical discriminative models, to detect abnormal behavior in LOB time series.

This paper first and foremost fits in the financial market anomalies literature. Although research on machine learning–based anomaly detection is plentiful, literature on machine learning for financial market anomalies is relatively scarce, and most deep learning methods presented above have not been investigated in this context. This paper aims to reduce the gap between deep learning research and financial assets anomaly research. Ögüt et al. [64] were the first to investigate data mining algorithms (support vector machines and multilayer perceptron (MLP)) in the context of daily stock price manipulation detection. Using a labeled data set from January 1995 to March 2004, on an index traded on the Istanbul Stock Exchange, they find that data mining algorithms perform better in terms of total classification accuracy compared to multivariate statistical techniques (discriminant analysis and logistic regression). Their study utilizes daily data, which does not allow precise detection of the manipulations, meaning that regulators would have to analyze the complete daily trading data to find the fraudulent trades. But, because they use supervised learning algorithms, their methodology relies on labeled data. This is not desirable in practice, since it would require a substantial effort to generate a usable data set, and available fraud cases are very limited. This is the main factor driving the use of unsupervised models since they do not require any label. Furthermore, supervised models are only capable of detecting known patterns in the data, rendering them useless in unveiling emerging trade–based manipulations, or when market data's distribution inevitably drifts. These limitations support the adoption of unsupervised learning methods (Khodabandehlou and Golpayegani [26]).

Diaz et al. [16] were the first to introduce intraday trading data and an "open box" approach, in the form of decision trees, interpretable by market regulators to help detect trade–based manipulations. They find that the average traded volume is lower during manipulation periods, whereas liquidity, returns, and volatility are higher than usual. This corroborates with the empirical findings of Aggarwal and Wu [2]. Both studies relied on a limited set of trade–based manipulations enforced by the SEC that are unclear on the manipulation type, and their exact time of occurrence.

Cao et al. [6] propose a hidden markov model to detect certain trade–based manipulations: spoofing, pump-and-dump, and "others." From a set of four mid-price features, the proposed Adaptive Hidden Markov Model with Anomaly States (AHMMAS) can outperform standard machine learning (OC-SVM, kNN), and other statistical methods, on simulated anomalies injected in a LOB data set from the LOBSTER project (Huang and Polak [23]). This approach has some important drawbacks. The biggest one being the curse of dimensionality faced by the model. Indeed, two hidden states, one normal and one anomalous, are created for each feature, bringing the total number of hidden states to 2^n, for n the number of features. Given that the Viterbi algorithm (Forney [18]) is employed to determine the most probable series of hidden states in a time series of length T, the anomaly detection complexity is $\mathcal{O}(T \times 2^{2^n})$. We argue that more features, such as: market event interarrival time, cancelation volume, trade volume, LOB volume, LOB prices, etc., are necessary to capture a larger range of fraudulent patterns, which is problematic for AHMMAS. On the other hand, deep learning methods have shown great performance in high-dimensionality problems (Wang et al. [60]) and are thus better suited to generalize trade-fraud detection to multiple patterns. In a second paper, Cao et al. [7], the authors work directly with the orders, rather than the LOB, and use the OC-SVM algorithm to detect spoofing and quote stuffing orders from their volume, price, and duration, achieving state-of-the-art detection capabilities at the time, in terms of area under the receiver operating characteristic curve. But the temporal context of the orders is ignored, so that significantly large and/or rapid orders sent by high-frequency traders, or trades walking the book, may be wrongly classified as anomalous. We propose to include such context so that those types of orders can be safely ignored when pertinent, thus decreasing the false positive rate.

A series of studies are applying unsupervised hybrid models to first learn a representation of price–based time series features, and then cluster them to detect abnormal orders. Abbas et al. [1] use empirical mode decomposition (EMD), Close and Kashef [12]; Rizvi et al. [45] employ the dendritic cell algorithm, and Rizvi et al. [47] apply kernel–based principal component analysis. All of which then utilize KDE–based clustering techniques on the time series' representations, and all outperform the hidden markov model of Cao et al. [6]. The simulated fraudulent patterns in Abbas et al. [1]; Cao et al. [6, 7]; Close and Kashef [12]; Rizvi et al. [45, 47] are all fixed and repeated, thus probably leading to an overestimation of their methodologies' performances. Since we aim to build a more versatile and robust anomaly detection model, our set of simulated anomalies used out-of-sample includes more families of trade–based fraud, and they are also stochastically generated, leading to a greater range of fraudulent scenarios. In all these studies, only price features are used to detect potential frauds, which lacks generality to detect non-price related tactics, e.g., quote stuffing (Khodabandehlou and Golpayegani [26]). Furthermore, the chronological order of market events is important to characterize trade–based manipulations. For example, in pump-and-dump, there needs to be price ramping before observing large quantities of executions and cancelations. If that sequence of events is not respected, it is not fraudulent. It is unclear if the methods of Abbas et al. [1]; Close and Kashef [12]; Rizvi et al. [45, 47] are able to consider that chronological aspect. Alternatively, deep recurrent and attention models do consider the events' order in their time series representation.

Newer studies in financial market anomaly detection have started to focus on deep learning approaches. Leangarun et al. [28] use a supervised MLP to detect stock price manipulation using level 1 data. They can detect synthetic pump-and-dump cases with an accuracy above 88%, but without much success on synthetic spoofing cases. In a subsequent study, Leangarun et al. [29], employ a deep unsupervised framework based on GANs with LSTM network generators and discriminators utilizing uniformly sampled LOBs as input to achieve close to 70% accuracy out-of-sample on synthetic pump-and-dump activities. In a third paper, Leangarun et al. [30], the authors compare their LSTM-GAN to a LSTM-AE. They find that the AE model outperforms the GAN on synthetic pump-and-dumps, and both unsupervised models can detect five out of six real manipulation cases from the Stock Exchange of Thailand. Chullamonthon and Tangamchit [10] apply a supervised Transformer encoder model to detect both synthetic pump-and-dumps, and the same real fraud cases of Leangarun et al. [30]. The model achieves higher accuracy than the MLP of Leangarun et al. [28]. All these studies only focus on pump-and-pumps. Instead, we are proposing a generic framework able to detect different types of frauds from LOB data, a natural step in the next generation of market manipulation detectors (Khodabandehlou and Golpayegani [26]). Rizvi et al. [46] employ a MLP-AE trained on stock prices affinity matrices. The learned representations are then clustered with the kernel density estimation–based method proposed in Rizvi et al. [45]. This hybrid approach outperforms the works of Cao et al. [6] and Abbas et al. [1], cementing the idea that deep unsupervised models can learn better representations compared to previous methods. Again, it is unclear how the sequential aspect of trade–based manipulations is considered in this model, and the lack of non-price features is problematic for the detection of various trade–based manipulation types.

3. Methodology

3.1. Problem

We can formalize the problem as follows. Given an out-of-sample time series of length T and dimensionality m, $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T\},\ \mathbf{x}_t \in \mathbb{R}^m\ \forall t$, we are interested in finding subsequences of fixed length $k \ll T$, $S = \{S_t = \{\mathbf{x}_{t-k+1}, \mathbf{x}_{t-k+2}, \dots, \mathbf{x}_t\} \in \mathbb{R}^{k \times m} \mid t = k, \dots, T\}$, that strongly deviate from normalcy. Hence, we need to predict the series $Y = \{y_k, y_{k+1}, \dots, y_T\}$, where $y_t \in \{0 : \text{normal},\ 1 : \text{abnormal}\}$ is associated to S_t. In our context, the time series X contains LOB features useful in characterizing and detecting trade-–based manipulation patterns, which will be presented in Section 4.

3.2. Approach

Lately, the literature on anomaly detection is seeing a resurgence of interest in one-class classifiers (e.g., OC-SVM, Schölkopf et al. [51], and SVDD, Tax and Duin [56]), applied on representations learned by a higher-level deep neural network, called an encoder. The models are trained end-to-end and achieve state of the art results on many benchmarks (Ruff et al. [49]). Unfortunately, this approach also suffers from collapses into trivial solutions, which makes training difficult. Recently, Sohn et al. [54] have proposed a simple fix, where they suggest to first train an encoder network on some auxiliary task, and separately train the one-class classifier on these representations, as two separate learning steps, achieving impressive results on image benchmarks.

We propose to use the same approach, with some modifications for the financial domain. In their paper, Sohn et al. [54] focus on contrastive learning on augmentations (such as rotations, or flips) of the original images. Unfortunately, it is difficult to extend their approach to financial time series since possible notions of augmentations are less clear. Indeed, because the semantics of LOB time series are very intricate, blindly applying previously proposed data augmentations breaks their nature, which makes contrastive learning methods difficult. Instead, we propose to train the encoder to solve the typical autoencoding task jointly with some decoder network on selected LOB features.[31] Note that other self-supervised or unsupervised tasks could be employed to train the encoder, such as masked language modeling, or some other prediction task. But masked-autoencoders have been found to have poor anomaly detection capabilities (Reiss et al. [44]), and predicting the stock market is arduous (Gandhmal and Kumar [19]), which drastically increases the complexity of learning efficient representations of LOB time series. For this reason, we prefer to adopt the autoencoding task in this context. After training, this decoder network is discarded, and a separate one-class classifier is trained, keeping the encoder fixed. We now detail each step of the proposed methodology.

[31] The autoencoding task is an unsupervised learning technique which consists of encoding an efficient lower–dimensional representation of the data and then decoding, i.e., reconstructing, the initial data from that representation.

3.3. Step 1: Autoencoding

In the first step, we train the deep neural network encoder $\phi_E : \mathbb{R}^{k\times m} \mapsto \mathbb{R}^d$ with parameters θ_E, jointly with another deep neural network $\phi_D : \mathbb{R}^d \mapsto \mathbb{R}^{k\times m}$ with parameters θ_D, called the decoder, where $d \ll k \times m$. The composition of the encoder and decoder networks forms the autoencoder, $\phi_D \circ \phi_E = \phi_{AE} : \mathbb{R}^{k\times m} \mapsto \mathbb{R}^{k\times m}$ with parameters $\Theta = \{\theta_E, \theta_D\}$. The goal of the encoder is to generate a representation vector, $\mathbf{z}_t \in \mathbb{R}^d$, semantically rich enough to summarize the input sequence $S_t \in \mathbb{R}^{k\times m}$, so the decoder is able to reconstruct that same sequence only from $\mathbf{z}_t$, i.e., $\phi_D\Big(\mathbf{z}_t = \phi_E(S_t \mid \theta_E) \Bigm| \theta_D\Big) = \phi_{AE}(S_t \mid \Theta) \approx S_t$. The proposed autoencoder is defined in this next part.

3.3.1. Bottlenecked Transformer autoencoder.

The architecture of the proposed autoencoder mostly follows the Transformer architecture of Vaswani et al. [57], along with some modifications. Figure 19 details the original architecture (left) and the proposed model (right).

Initially proposed in the field of natural language processing, Transformers are deep learning models adopting the encoder-decoder architecture of earlier sequence-to-sequence models. Transformers process entire sequences at once, instead of relying on recursion like deep recurrent models. They do so using self-attention in the "Multi-Head Attention" module (see Figure 19), allowing them to focus, i.e., put more weight, on relevant portions of the input data (timewise and feature-wise) depending on the sequence itself, to create more informative sequence representations. The Transformer encoder consists of N encoding layers, each successively transforming the sequence representation into a final matrix containing the sequence's contextual information called the encoding. The Transformer decoder is then tasked to autoregressively generate an output sequence based on this encoding, without attending the current or future observations of the expected output sequence. We refer the reader to Vaswani et al. [57] for its complete description, as we want to focus on the alterations we make to the model.

As can be seen in Figure 19, the differences with the architecture of Vaswani et al. [57] are the addition of a bottleneck composed of the "flatten" and "linear" blocks, and

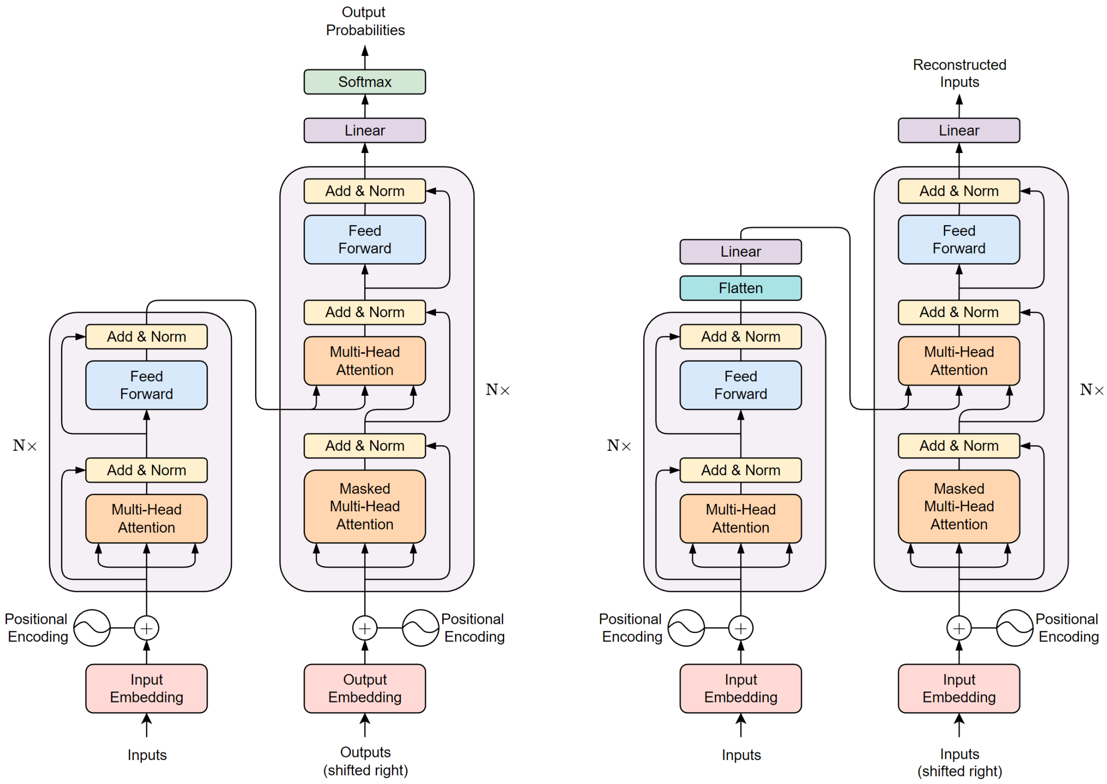

Fig. 19. Original Transformer autoencoder of Vaswani et al. [57] (left) and the proposed bottlenecked Transformer autoencoder (right). The encoder representation matrix in $\mathbb{R}^{k\times d}$ is flattened to a vector in $\mathbb{R}^{kd}$, then passed to a linear layer to reduce the representation further to $\mathbb{R}^{d}$, before being fed to the decoder.

the removal of the "softmax" layer at the end of the decoder. Similar models have been proposed for sentence embedding (Montero et al. [36]; Wang et al. [60]), and musical style encoding (Choi et al. [9]), but never in the context of anomaly detection. The bottleneck deserves some additional details. After N Transformer-encoder layers, the input matrix $S_t = (\mathbf{x}_{t-k+1}, \ldots, \mathbf{x}_t) \in \mathbb{R}^{k\times m}$ is mapped to a matrix $E_t = (\mathbf{e}_{t1}, \ldots, \mathbf{e}_{tk}) \in \mathbb{R}^{k\times d}$ following the standard Transformer architecture of Vaswani et al. [57]. After that, the bottleneck flattens, i.e., concatenates, this representation matrix into a kd-vector $\mathbf{e}_t$:

$$\mathbf{e}_t = \begin{bmatrix} \mathbf{e}_{t1} & \ldots & \mathbf{e}_{tk} \end{bmatrix} \in \mathbb{R}^{kd},$$

which is then linearly projected into a d-vector, yielding the representation:

$$\mathbf{z}_t = W\mathbf{e}_t + \mathbf{b} = \phi_E(S_t \mid \theta_E) \in \mathbb{R}^{d},$$

where $W \in \mathbb{R}^{d \times kd}$ and $\mathbf{b} \in \mathbb{R}^{d}$. Because $d \ll k \times m$, information is lost in this bottleneck. In turn, this means the decoder only attends to a limited subset of information summarizing the input sequence, and the encoder is forced to learn a semantically rich temporal context vector $\mathbf{z}_t$.

3.3.2. Autoencoder training.

The autoencoder ϕ_{AE} is trained to minimize the L_2 reconstruction loss of any subsequence S_t, which is the standard autoencoding task:

$$\mathcal{L}(\Theta) = \mathbb{E}\Big[\Big(S_t - \phi_{AE}(S_t \mid \Theta)\Big)^2\Big],$$
$$\Theta^* = \arg\min\nolimits_{\Theta} \mathcal{L}(\Theta).$$

Hence, supposing a fraud–free data set of N subsequences $D = \{s_1, \ldots, s_N\},\ s_i \in \mathbb{R}^{k \times m}$, the parameters of ϕ_{AE} are estimated by stochastic gradient descent on the empirical measure:

$$\widehat{\Theta} = \arg\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} \Big|\Big| s_i - \phi_{AE}(s_i \mid \Theta) \Big|\Big|_{\mathrm{F}},$$

for $|| \cdot ||_{\mathrm{F}}$ the Frobenius norm.

3.4. Step 2: Dissimilarity learning

In the second step, we discard the decoder ϕ_D, and train a OC-SVM on the representations $\mathbf{z}_i = \phi_E(s_i \mid \widehat{\theta}_E),\ i = 1, \ldots, N$, learned by the encoder on the fraud-free set D. The OC-SVM of Schölkopf et al. [51] is a novelty detection algorithm extending the support vector machine algorithm (SVM, Cortes and Vapnik [13]) to the unsupervised case. It finds a subset of the input space such that a new point drawn from the same distribution as the data will lie outside the subset with arbitrarily small probability, leaving abnormal data points outside, thus allowing us to detect any anomalous encoder representations. We introduce the algorithm here for completeness.

Defining the feature map $\Phi : \mathbb{R}^d \mapsto \mathcal{F}$ such that $\mathcal{F}$ is a space where the dot product in the image of Φ can be obtained by a kernel function, i.e.:

$$k(\mathbf{z}, \mathbf{z}') = \Phi(\mathbf{z}) \cdot \Phi(\mathbf{z}'),$$

the OC-SVM tries to separate the data from the origin in $\mathcal{F}$ with the hyperplane $\Big\{\Phi(\mathbf{z}) \Bigm| \mathbf{w} \cdot \Phi(\mathbf{z}) = \rho\Big\}$ with maximum margin. If no such hyperplane exists, slack variables are added to some mapped representations (the support vectors), while also allowing points in the origin's half space (the outliers). That is, the OC-SVM solves the quadratic program:

$$\min_{\boldsymbol{w}\in\mathcal{F},\boldsymbol{\xi}\in\mathbb{R}^N,\rho\in\mathbb{R}} \frac{1}{2}||\mathbf{w}||^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho$$
$$\text{s.t. } \mathbf{w}\cdot\Phi(\mathbf{z}_i) \geq \rho - \xi_i,\ \xi_i \geq 0, \quad \forall i = 1,\ldots,N,$$

where the hyperparameter $\nu \in (0,\, 1)$ controls the upper bound on the fraction of outliers, and ξ_is are the slack variables. We use the radial basis function:

$$k_\gamma(\mathbf{z},\mathbf{z}') = \exp(-\gamma\|\mathbf{z}-\mathbf{z}'\|^2)$$

as the kernel with hyperparameter $\gamma > 0$. The optimization problem can be solved efficiently with any standard quadratic programming solver. The OC-SVM decision function has a solution of the form:

$$f(\mathbf{z}) = \operatorname{sign}\left(\sum_{i=1}^{N}\widehat{\alpha}_i k_\gamma(\mathbf{z},\mathbf{z}_i) - \widehat{\rho}\right) \in \{-1{,}1\},$$

where outliers have a negative value. The $\widehat{\alpha}_i$ and $\widehat{\rho}$ estimated on the representations $\mathbf{z}_i$ are kept for the dissimilarity function detailed in the next subsection.

3.5. Step 3: Predicting

In the third step, once both the encoder and the OC-SVM are trained, one makes a pass over the out-of-sample subsequences of set S, and record their dissimilarity value, which is a slight modification of the OC-SVM decision function:

$$\text{dissimilarity}(S_t) = \widehat{\rho} - \sum_{i=1}^{N}\widehat{\alpha}_i k_\gamma\Big(\phi_E\big(S_t \mid \widehat{\theta}_E\big), \phi_E\big(s_i \mid \widehat{\theta}_E\big)\Big) \in \mathbb{R},\ \forall S_t \in S.$$

This function quantifies how far the representation of the subsequence S_t is from the support learned by the OC-SVM. In other words, the greater the dissimilarity, the further the subsequence is from normal data (and the hyperplane learned by the OC-SVM), and the more anomalous it is. Finally, we classify a subsequence S_t as anomalous if

$$\text{dissimilarity}(S_t) > \tau,$$

for some threshold $\tau \in \mathbb{R}$ controlling the sensitivity of the algorithm, i.e.,

$$\widehat{y}_t = \mathbb{I}_{\{\text{dissimilarity}(S_t)>\tau\}}, \, \forall S_t \in S,$$

for $\mathbb{I}_{\{\cdot\}}$ the indicator function. Previous financial anomaly detection papers (e.g., Leangarun et al. [30]) have used the L_2 reconstruction loss of autoencoders for this task, $\left\| S_t - \phi_{AE}(S_t \mid \widehat{\Theta}) \right\|_{\mathrm{F}}$, but as we will show, this leads to suboptimal detection performance on our data set. Indeed, the reconstruction loss is highly volatile in time so true anomalies are difficult to detect, generating a high false positive ratio. Our approach instead utilizes the dissimilarity function, dissimilarity(S_t), to quantify the abnormality of S_t, yielding better results because normal data is more easily discernible in the representation space. These aspects will be discussed in detail in Section 5. Deep learning models have not been well explored for collective anomalies, unlike point anomalies (Pang et al. [40]), and therefore our paper also contributes to fill this gap in the time series literature with a direct application to LOB data. Figure 20 presents the overall proposed anomaly detection model combining the bottlenecked Transformer encoder, and the OC-SVM algorithm.

4. Data

As explained in Section 3, our approach is unsupervised, which means that we do not need explicit examples of fraud to train and run our model. However, it is useful to have such examples to evaluate our algorithm and compare it to alternative approaches out-of-sample. To do so, we will use the standard strategy of using a fraud–free data set of orders, and artificially add in examples of fraud in the data set, given the scarcity of actual fraudulent

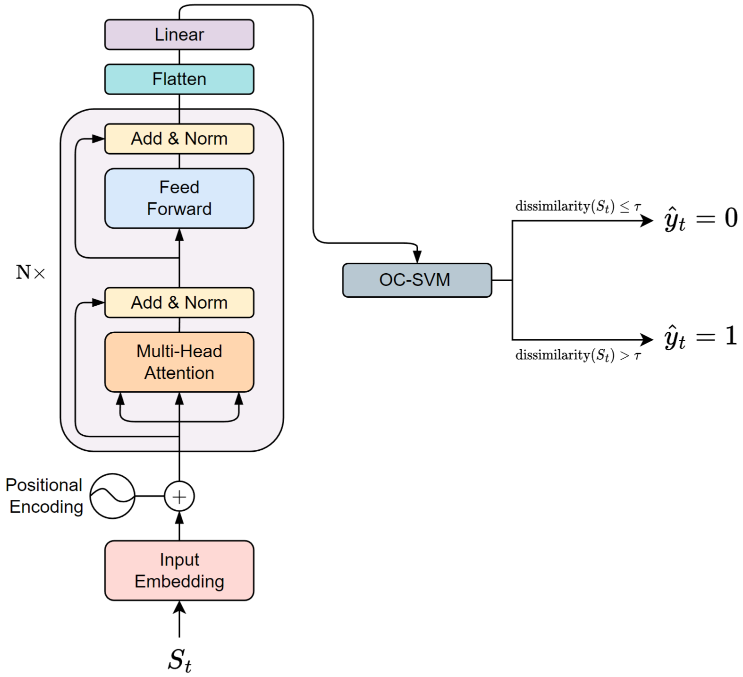

Fig. 20. Proposed dissimilarity–based anomaly detection framework.

events. Note that although these examples of fraud will only be present in the test data, our approach will not be aware of their location, nor amount, so that the methodology remains unsupervised. We now describe the data in further details.

4.1. Base data set

Our base data comes from the LOBSTER project (Huang and Polak [23]). It contains the LOB level 1 (L1) data of five NASDAQ stocks: Amazon (ticker: AMZN), Apple (ticker: AAPL), Google (ticker: GOOG), Intel (ticker: INTL), and Microsoft (ticker: MSFT) on June 21, 2012. The LOBSTER project data has also been used in most financial market fraud detection papers (Abbas et al. [1]; Cao et al. [6, 7]; Close and Kashef [12]; Leangarun et al. [28]; Rizvi et al. [45, 46, 47]), so it is a good point of comparison. The data is split into two files, one for market events, and one for rebuilt LOBs. Table 38 details the data, and Table 39 provides some descriptive stock statistics. It is from that raw data that LOB features are created to be used in the model of Section 3.

Table 38. Description of the information contained in LOBSTER project's messages and LOBs files.

Messages		LOBs	
Variable	**Description**	**Variable**	**Description**
Time	Nanoseconds past midnight	Best bid price	Best buying price
Type	1: Submission new limit order	Best bid size	Total number shares available at bid price
	2: Partial cancelation limit order	Best ask price	Best selling price
	3: Deletion limit order	Best ask size	Total number shares available at ask price
	4: Execution visible limit order		
	5: Execution hidden limit order		
OrderID	Unique order reference number		
Size	Number shares of order		
Price	Price of order		
Direction	-1: Sell limit order		
	1: Buy limit order		

Table 39. Descriptive statistics of LOBSTER stocks on June 21, 2012.

Statistic | Stock	AAPL	AMZN	GOOG	INTC	MSFT
Quotes	54,818	27,845	24,368	202,231	205,695
Trades	34,990	11,419	11,678	32,483	33,414
Cancels/Quotes	52.34%	65.54%	55.14%	84.20%	83.76%
Std. Mid-price	2.99	1.36	4.28	0.27	0.29
Avg. Bid-Ask Spread (bps)	2.66	6.10	5.45	4.92	4.40
Avg. Best Bid Size	227.20	249.34	172.11	17,194.81	14,965.07
Avg. Best Ask Size	147.19	145.36	134.99	14,360.18	15,796.58
Avg. Order Size	88.30	95.90	81.89	503.46	600.43
Avg. Trade Size	81.46	71.00	60.47	322.37	323.95

4.2. Synthetic manipulations

As mentioned earlier, reported trade–based manipulation cases are rare, hence most studies have instead simulated some, and inserted them back in their initial fraud–free data set. We also follow that methodology, but we significantly increase the complexity of fraudulent patterns, as explained in Section 2. Not only do we simulate more families of trade frauds, but we also stochastically generate them so that multiple scenarios are included in the study, while making sure not to denature the manipulations. Previous papers repeat the same fixed, limited, set of orders, which probably results in an overestimation of their model's true detection potential. Three distinct trade–based manipulation types are studied: pump-and-dump, layering, and quote stuffing. 50 trade–based manipulations of each type are randomly added per stock. 30 seconds of fraud–free data before and after each manipulation are kept to create the test set containing all frauds, so that the vast majority of the test data is also fraud–free, keeping a low anomaly ratio. The remainder of the LOB data is split to form the training set, representing 70% of the remainder, and the validation set, representing the last 30%. Given the large discrepancy in the data quantity between Amazon, Apple, Google and Intel, Microsoft (see Table 39), we oversample the first group of stocks in the training and validation data sets to get a more balanced representation. The training set is used to train the bottlenecked Transformer autoencoder with the L_2 reconstruction loss detailed in Section 3.3.2, and the validation set is needed to select the optimal parameters $\widehat{\Theta}$ generating the lowest loss on that set. Then, the training and validation sets are concatenated to form a single fraud–free set on which the OC-SVM of Section 3.4 is trained to learn the dissimilarity function. The test set is only utilized to evaluate the methodology's out-of-sample performance. We now describe in detail the simulated trade–based manipulations.

4.2.1. Pump-and-dump.

The U.S. Securities and Exchange Commission (SEC) defines pump-and-dumps as: "[...] schemes [that] have two parts. In the first, promoters try to boost the price of a stock [...]. Once the stock price has been pumped up, fraudsters move on to the second part, where they seek to profit by selling their own holdings of the stock, dumping shares into the market."[32]

We closely follow the procedure of Chullamonthon and Tangamchit [11] to simulate pump-and-dump patterns. We replicate their "low degree" pump-and-dump scenario in which a price change of 3–4% occurs, with an increase in both bid and ask volumes of 25–100% during a pump period of one second. Following that price pump, a dumping stage of three seconds sees the same increased level of canceled bid volume and matched volume, bringing the LOB prices to their initial level. They also propose two other, more aggressive scenarios, but we only replicate the lowest degree of fraud to remain more conservative. The different scenarios of Chullamonthon and Tangamchit [11] are based on cryptocurrencies pump-and-dumps found in Kamps and Kleinberg [25], and on an event identified by Nanex Research, a Chicago–based firm specializing in high-frequency trading data; The Westinghouse Air Brake

[32] https://www.investor.gov/protect-your-investments/fraud/types-fraud/pump-and-dump-schemes (accessed May 23, 2023).

Technologies Corp stock's (NYSE:WAB) price jumped 8% in the span of one second, and then dropped back close to its initial value after three seconds on December 14, 2011 (Nanex Research [38]). Table 40 lists the pump-and-dump characteristics and the ranges used to randomly simulate the manipulations, and Figure 21 presents a simulated pump-and-dump pattern.

Table 40. Range of randomly simulated pump-and-dump scenarios.

Characteristic	Total Price Increase	Non Bona Fide Order Size	Pump Duration	Dump Duration
Range	[3, 4]%	[1.25, 2]× Average L1 Volume	[750, 1,250] ms	[2,250, 3,750] ms

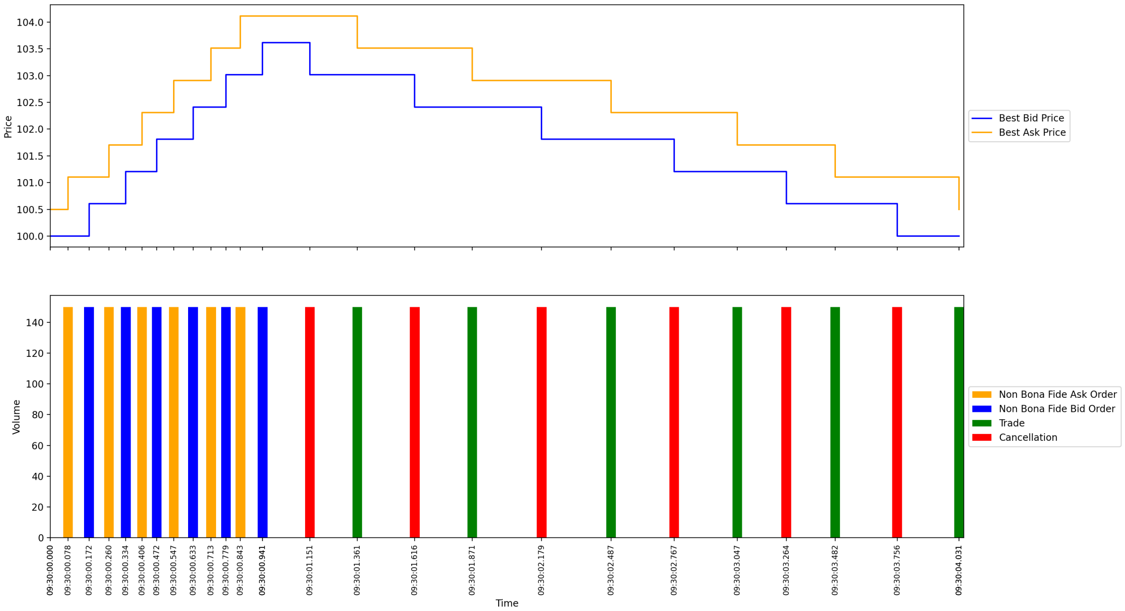

Fig. 21. Toy example of "low degree" pump-and-dump manipulation.

4.2.2. Layering.

The Investment Industry Regulatory Organization of Canada (IIROC) defines layering as: "[...] [the act of] placing a *bona fide* order on one side of the market while simultaneously "layering" orders in the consolidated market display on the other side of the market without intention to trade [...] as [to induce] a false or misleading appearance of trading activity or artificial price. In this case, the purpose of the "layering" is to "bait" other market participants to react and trade with the *bona fide* order on the other side of the market at an artificial price." (IIROC [24])

On April 30, 2014, the SEC reported layering activities that occurred on the stock W.W. Grainger (NYSE: GWW) on June 4, 2010.[33] Figure 22 details the layering orders listed in the official document. At 11:08:55.152, the trader placed a *bona fide* order to sell 1,000 units of the stock at $101.34 per share when the best bid was at $101.27 and the best ask at $101.37. From 11:08:55.164 to 11:08:55.323, the trader sent 11 orders to buy GWW at increasing prices, totaling a volume of 2,600 shares, pushing the best bid price to $101.33. Another market participant, deceived by the layering orders, traded against the *bona fide* sell order of 1,000 shares at 11:08:55.333. At 11:08:55.932, the trader canceled all the *non bona fide* bid orders, and the L1 prices returned to their initial value.

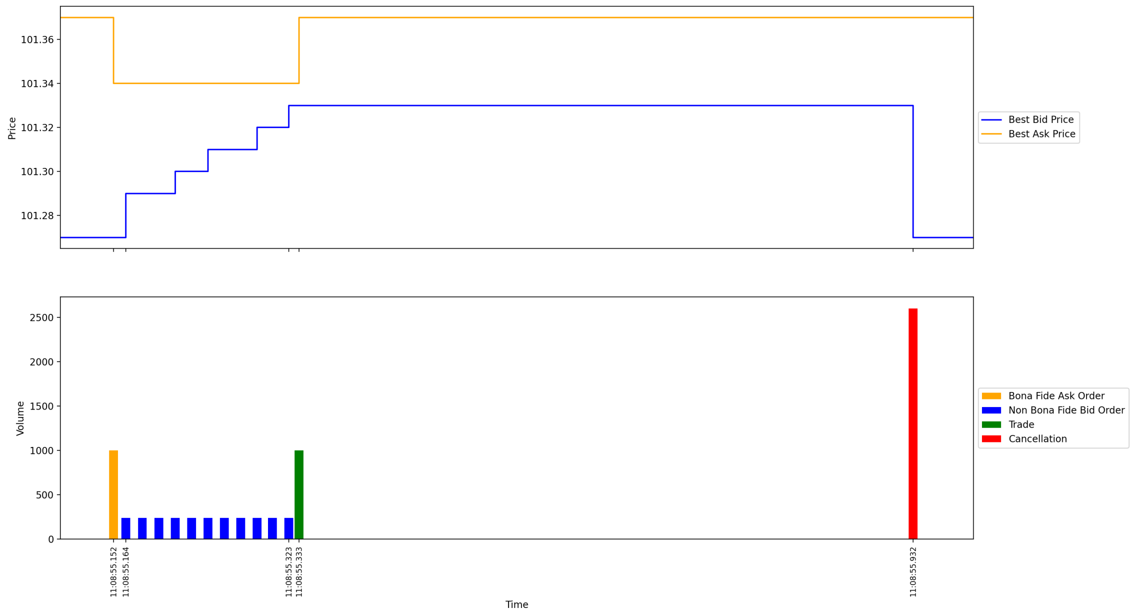

Fig. 22. Recreation of the layering activities found by the SEC in NYSE:GWW on June 4, 2010.

From this case, multiple stylized facts of interest can be used to replicate layering patterns: the *bona fide* order is placed around 3 bps inside the bid-ask spread, a sequence of 11 *non bona fide* orders are placed at a rate of one per 14 ms, the *non bona fide* orders push the

[33] https://nj.gov/oag/newsreleases14/Hold-and-Tobias-Consent-Order-05-02-14.pdf (accessed March 13, 2023).

price around 6 bps away from its initial value, a trade occurs rapidly after a *non bona fide* order and their placement are stopped, all the *non bona fide* orders are canceled less than 600 ms after the trade, and finally, the *bona fide* order is 2.6 times smaller than the total volume of the *non bona fide* orders. Another important stylized fact is that spoofing orders are on average 5.6 times larger than typical orders (Lee et al. [31]).[34] Table 41 lists layering characteristics and the ranges used to randomly simulate the manipulations.

Table 41. Range of randomly simulated layering scenarios.

Characteristic	**Bona Fide Order Side**	**Bona Fide Order Price**	**Bona Fide Order Size**	**Nb. Non Bona Fide Orders**	**Non Bona Fide Interarrival Time**
Range	$\{Bid, Ask\}$	$Bidprice + (0,3)$ *bps* if Bona Fide Side = Bid $Askprice - (0,3)$ *bps* if Bona Fide Side = Ask	[2, 3]× Non Bona Fide Total Size	$[10, 12]$	$[10, 20]$ *ms*
Characteristic	**Non Bona Fide Total Size**	**Non Bona Fide Price Movement**	**Trade Delay**	**Cancelation Delay**	
Range	[5, 6]× Average order size	$(0, 6]$ *bps*	[5, 15] *ms*	$[100, 1100]$ *ms*	

4.2.3. Quote stuffing.

IIROC defines quote stuffing as: "[...] the input by a Participant or Access Person of excessive market data messages with the intent to "flood" systems and create "information arbitrage" opportunities for itself [...]." (IIROC [24]).

Egginton et al. [17] find that quote stuffing occurs often on only one side of the book at a time. They also notice a drastic increase in the new order and cancelation rates, a decrease in order size, and an augmentation of order updates slightly inside the spread during quote stuffing periods. Nanex also discovered multiple quote stuffing algorithm imprints. More specifically, they detail the orders sent by Citadel Securities for which a disciplinary action was taken by NASDAQ in 2014.[35] For example, from 13:32:53.029 to 13:33:00.998, Citadel placed 8 to 9 orders to buy 100 shares of NASDAQ:PENN, per millisecond, before immediately canceling them. This caused delays up to 16 ms in the U.S. Securities Information Processor, creating arbitrage opportunities (Nanex Research [39]). Nanex has also found that order rates at 10 per millisecond and above will create delays in NYSE's consolidated quotation system (Nanex Research [37]). Although quote stuffing sequences can last thousands of events, we limit their length as to not have a disproportionate number of fraudulent

[34] Multiple spoofing orders in a sequence generate a layering manipulation.

[35] NASDAQ Disciplinary Actions against Citadel, accessed March 15, 2023

orders in our data set. Considering these stylized facts, Table 42 lists quote stuffing characteristics and the ranges used to randomly create the manipulations. Figure 23 shows a toy example of quote stuffing activity where a high-frequency trader submits, then almost immediately cancels, limit orders inside the bid-ask spread. In the span of six milliseconds, 25 buy limit orders are sent to the market then rapidly canceled, thus sending a total of 50 messages to other market participants.

Table 42. Range of randomly simulated quote stuffing scenarios.

Characteristic	**Fraud Side**	**Nb. Events**	**Order Rate**	**Order Size**	**Order Price**
Range	$\{Bid, Ask\}$	$[50, 200]$	[8, 10] / ms	[1, 10]th percentile order volume distribution	$Bidprice + (0,\, midprice)$ if Fraud Side $= Bid$ $Askprice - (0,\, midprice)$ if Fraud Side $= Ask$

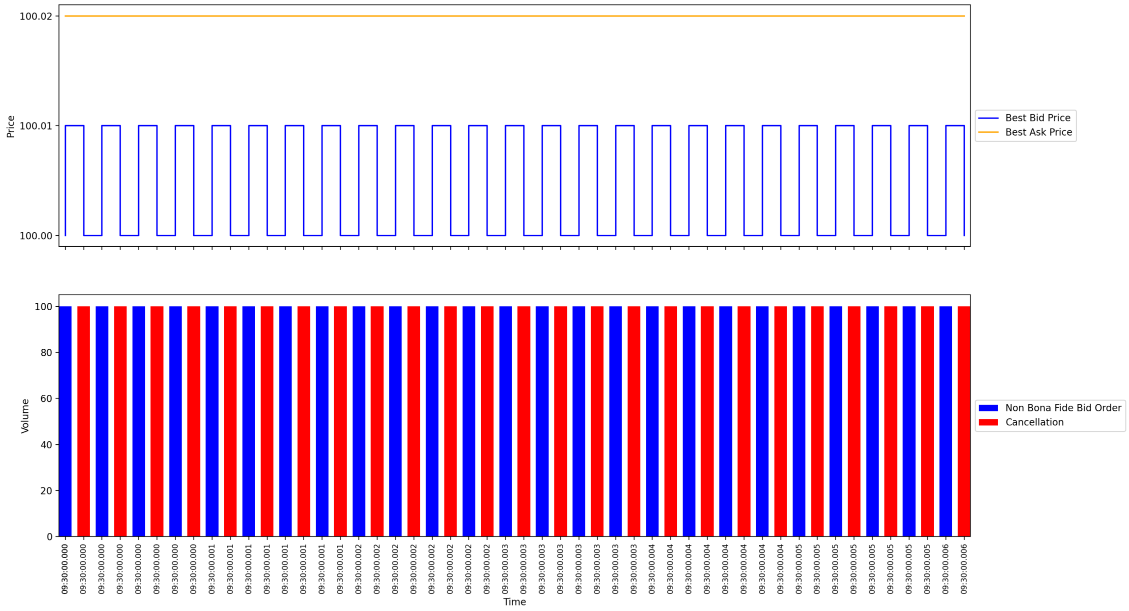

Fig. 23. Toy example of quote stuffing.

4.3. Feature engineering

As mentioned previously, the proposed framework utilizes a set of features aiming to best describe the variations in LOB states so that multiple manipulation types can be detected from a single model. The trade–based manipulations described above are all very distinct, and an adequate set of features needs to be able to describe them all. Thus, we propose LOB–based features that cover price returns, volumes, cancelation volumes, and trade volumes, while also considering the rapidity at which the different market events occur. Table 43 details all the features used in the framework built from the original data presented at the beginning of this section. All 14 features are standardized on a daily stock-per-stock basis. For any daily stock feature vector $\mathbf{x}$, its standardized version, $\mathbf{x}'$, is given by:

$$\mathbf{x}' = \frac{\mathbf{x} - \overline{x}}{\sigma},$$

where $\overline{x}$ and σ are its empirical mean and standard deviation, respectively. In Table 43, $P_t^{Bid/Ask}$ are the best bid/ask prices of the LOB at market event t, $Size_t^{Bid/Ask}$ the best bid/ask size of the LOB, $Size_{Cancel/Trade,t}^{Bid/Ask}$ the size of the cancelation/trade on the best bid/ask side, and Δt the time delta between market events t and $t-1$.

Table 43. List of LOB features proposed to capture trade–based manipulations.

Feature	Description
r_t^{Bid}	Best bid-price return at event t: $r_t^{Bid} = \ln\left(P_t^{Bid}/P_{t-1}^{Bid}\right)$
r_t^{Ask}	Best ask-price return at event t: $r_t^{Ask} = \ln\left(P_t^{Ask}/P_{t-1}^{Ask}\right)$
$r_t^{Bid}/\Delta t$	Difference quotient of best bid-price return *w.r.t.* time, at event t
$r_t^{Ask}/\Delta t$	Difference quotient of best ask-price return *w.r.t.* time, at event t
$\overline{Size}_t^{Bid}$	Simple moving average of total size at best bid-price, at event t: $\overline{Size}_t^{Bid} = \sum_{k=0}^{9} Size_{t-k}^{Bid}/10$
$\overline{Size}_t^{Ask}$	Simple moving average of total size at best ask-price, at event t: $\overline{Size}_t^{Ask} = \sum_{k=0}^{9} Size_{t-k}^{Ask}/10$
$Size_{Trade,t}^{Bid}$	Size of trade consuming liquidity at best bid-price, at event t
$Size_{Trade,t}^{Ask}$	Size of trade consuming liquidity at best ask-price, at event t
$Size_{Cancel,t}^{Bid}$	Size of order cancelation/deletion located at best bid-price, at event t
$Size_{Cancel,t}^{Ask}$	Size of order cancelation/deletion located at best ask-price, at event t
$I_{Trade,t}^{Bid}$	Indicator of trade rapidity on best bid-price, at event t: $I_{Trade,t}^{Bid} = \mathbb{I}_{\{Size_{Trade,t}^{Bid} \neq 0\}}/\Delta t$
$I_{Trade,t}^{Ask}$	Indicator of trade rapidity on best ask-price, at event t: $I_{Trade,t}^{Ask} = \mathbb{I}_{\{Size_{Trade,t}^{Ask} \neq 0\}}/\Delta t$
$I_{Cancel,t}^{Bid}$	Indicator of cancelation rapidity on best bid-price, at event t: $I_{Cancel,t}^{Bid} = \mathbb{I}_{\{Size_{Cancel,t}^{Bid} \neq 0\}}/\Delta t$
$I_{Cancel,t}^{Ask}$	Indicator of cancelation rapidity on best ask-price, at event t: $I_{Cancel,t}^{Ask} = \mathbb{I}_{\{Size_{Cancel,t}^{Ask} \neq 0\}}/\Delta t$

5. Experiments

5.1. Setup

The models are fitted on the training data set containing the features computed from the LOBSTER data of Huang and Polak [23] presented in Section 4, for every stock. They are trained for 250 epochs, with a batch size of 512, and on subsequences of length $k = 25$. Autoencoders have a representation dimension of $d = 128$, which is enough to compress the subsequence dimensionality of $k \times m = 25 \times 14 = 350$, while ensuring they are able to reconstruct the time series input sufficiently well. The rest of the bottlenecked Transformer autoencoder's hyperparameters are set similarly to Vaswani et al. [57]: $h = 8$, $d_{ff} = 4d = 512$, and $N = 6$, and the model is trained following their proposed methodology based on the Adam optimizer of Kingma et al. [27]. As for the OC-SVM, γ is set like in Sohn et al. [54], i.e., $\gamma = 10/\Big(d \times \mathrm{Var}\left(\phi_E(s_i \mid \widehat{\theta}_E)\right)\Big)$, and the other hyperparameters are set to their default values in scikit-learn (Pedregosa et al. [42]). No hyperparameter tuning of the autoencoder, nor the OC-SVM, is done as to avoid overfitting problems.

All metrics used in this section rely on combinations of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Most studies in the trade–based manipulation detection literature do not formally define them, and given the problem's context, their exact definitions are not clear. We formally introduce them to eliminate any ambiguity:

- TP: market events part of a fraud, and part of at least one detected subsequence,
- TN: market events not part of a fraud, and not part of any detected subsequences,
- FP: market events not part of a fraud, and part of at least one detected subsequence,
- FN: market events part of a fraud, and not part of any detected subsequences,

all of which are function of the dissimilarity threshold τ. Also, whenever a market event part of a fraud is detected, then we consider that all market events in that fraud are also detected. This is reasonable, since market regulators would catch the fraudulent behavior by looking at the surrounding events of a flagged anomaly.

As mentioned in Golmohammadi and Zaine [20], and Khodabandehlou and Golpayegani [26], the misclassification cost of fraudulent orders is higher than the one of normal orders, meaning that an emphasis on recall, as opposed to precision, is necessary to correctly evaluate trade–based manipulation detection frameworks. Hence, a $\beta > 1$ in F_β-measures is more appropriate. We detail the different statistics for any given τ:

$$Precision(\tau) = \frac{TP(\tau)}{TP(\tau) + FP(\tau)} \in [0{,}1]$$

$$Recall(\tau) = TPR(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)} \in [0{,}1]$$

$$F_\beta(\tau) = (1 + \beta^2) \cdot \frac{Precision(\tau) \cdot Recall(\tau)}{(\beta^2 \cdot Precision(\tau)) + Recall(\tau)} \in [0{,}1]$$

$$AUPRC = \int_{-\infty}^{\infty} Precision(\tau)Recall'(\tau)d\tau \in [0{,}1]$$

$$FPR(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)} \in [0{,}1]$$

$$AUROC = \int_{-\infty}^{\infty} TPR(\tau)FPR'(\tau)d\tau \in [0{,}1]$$

5.2. Main results

5.2.1. General performance and comparative study.

Like Golmohammadi and Zaine [20], we use the F_4-measure to quantitatively compare the different models, and we also provide the area under the precision-recall curve (AUPRC) to evaluate the models' general performance. The area under the receiver operating characteristic curve (AUROC) is also included as a third metric, since it is still used in recent literature, e.g., Close and Kashef [12]; Rizvi et al. [46]; Wang et al. [59], even though it can be misleading in imbalanced data sets, which is the case for anomaly detection (Saito and Rehmsmeier [50]). Table 44 compares the performance of our proposed method to some others put forward in past studies. Only unsupervised frameworks working on tick data, i.e., orders or LOBs, are considered, and all are trained and tested on our data. The optimal threshold τ^* for each method is selected to optimize the F_4-measure, and we also detail the precision and recall at that point of the precision-recall curve.

Table 44. Comparative study of the proposed method with previous unsupervised papers, on our data. Best metric in bold, second best is underlined.

Paper	Model	AUROC	AUPRC	F_4	Precision	Recall
Ours	Transformer-AE + OC-SVM	<u>0.900</u>	**0.847**	**0.935**	0.628	0.965
Abbas et al. [1]	EMD + KDE Clustering	0.700	0.375	0.803	0.195	0.992
Cao et al. [7]	OC-SVM	**0.944**	<u>0.732</u>	<u>0.899</u>	0.397	0.976
Leangarun et al. [29]	LSTM-GAN	0.694	0.323	0.774	0.197	0.878
Leangarun et al. [30]	LSTM-AE	0.682	0.293	0.773	0.167	1.000
Rizvi et al. [46]	MLP-AE + MKDE Clustering	0.674	0.371	0.776	0.169	1.000

By working on uniform sampling of the LOB with one second intervals, Leangarun et al. [29, 30] miss the entire temporal context of fast trade–based manipulations, like quote stuffing and layering. Furthermore, the impact of layering is minimal when looked at a second resolution, making it nearly impossible to detect in their methodology (see Section 4.2.2). This is represented in Table 44, where they achieve the lowest AUPRC and F_4-measure on our data set. The hybrid clustering techniques of Abbas et al. [1] and Rizvi et al. [46] fare a bit better by working directly on the LOB. But by only focusing on price features, they lack the LOB volume and cancelation features needed to detect various trade–based manipulations. Also, the temporal context is not considered in their methodology, which can be appropriate for point anomalies, but is not when trying to capture collective time series anomalies. This aspect is also a shortcoming of Cao et al. [7], where single orders are classified without any context. But still, they achieve a greater performance than newer methods because of better–crafted features able to encapsulate important orders' characteristics: size, LOB price effect, and duration. The dissimilarity objective of the OC-SVM algorithm also seems to be more apt than clustering (in Abbas et al. [1]; Rizvi et al. [46]), and estimation models (in Leangarun et al. [29, 30]), for trade–based fraud detection.

Overall, when tested on pragmatic simulations of sophisticated, and distinct, trade–based manipulations, all previously proposed unsupervised methods fall short when compared to our model in terms of AUPRC and F_4-measure, mostly because of their lower precision, i.e., higher false positive rate. For market regulators, this means that less time is wasted on false alarms, which is the main goal of this paper. The work of Cao et al. [7] is a close second, and also significantly outperforms the other, more recent models evaluated in Table 44. Finally, our methodology integrates more relevant features of the LOB, and their temporal context. It also learns a more descriptive representation of the data, enabling an easier separation of normal and abnormal subsequences compared to previous methods, which is corroborated by Table 44.

5.2.2. Performance per stock.

Table 45 details the proposed method's performance on a per-stock basis with the same F_4-optimal threshold τ^* found for Table 44. An important observation can be made from Table 45, which is that the model's performance is not drastically different for any stock in terms of F_4-measure, meaning that it was able to learn an asset–independent representation of normality. This is important for market regulators since a single model is enough to capture anomalies on all assets. This contrasts with rule–based systems where their rules have to be manually adjusted for each asset, depending on their stylized facts (see Table 39 for an overview of important stylized facts in stocks). Moreover, the F_4-measures are all approaching 1, demonstrating the great general performance of the methodology.

Table 45. Per-stock performance of the proposed methodology with the general F_4-optimal τ^*.

Stock	**F$_4$**	**Precision**	**Recall**
AAPL	0.955	0.710	0.977
AMZN	0.893	0.783	0.901
GOOG	0.940	0.820	0.949
INTC	0.949	0.530	0.998
MSFT	0.934	0.499	0.988

5.2.3. Performance per manipulation type.

For the first time in the financial market fraud detection literature, we quantitatively study the difficulty of detecting different trade–based manipulation types. Figure 24 uses the t-distributed stochastic neighbor embedding (t-SNE) of Maaten and Hinton [34] to visualize the representation space of dimension $d = 128$.[36] As can be observed, different clusters for each manipulation tactics lie outside the main cluster of normal data, meaning that the representation vectors are able to discriminate between normal and abnormal subsequences, even differentiating between fraud types. Figure 25 shows the empirical distribution of our trained dissimilarity measure on normal and fraudulent subsequences on all stocks, from the representation vectors shown in Figure 24. Visually, pump-and-dumps seem to be the easiest to detect given their smaller overlap with the distribution of normal data, whereas layering activities appear to be the hardest to detect. Quote stuffing resides in between the other two manipulation types.

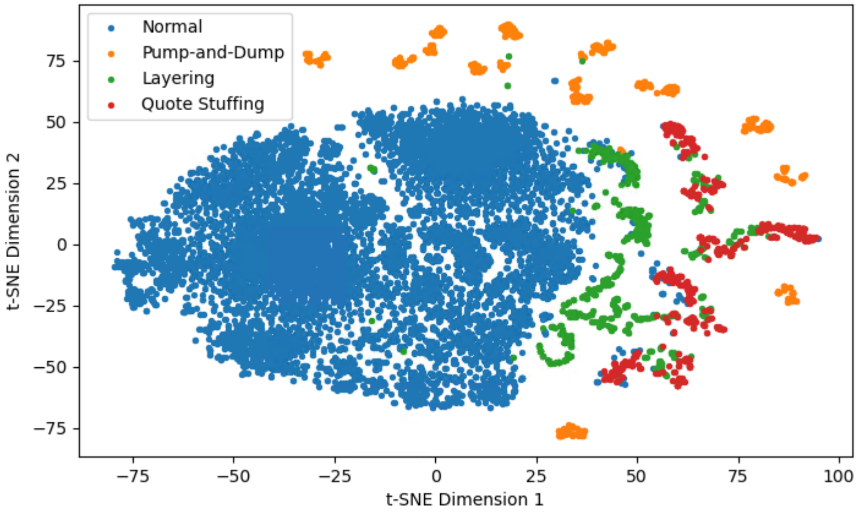

Fig. 24. t-SNE visualization of the representation space generated by the bottlenecked Transformer encoder.

Table 46 statistically confirms that ordering, based on first-order stochastic dominance of the dissimilarity distributions. It applies the two-sample Kolmogorov-Smirnov test where

[36] t-SNE is a nonparametric, nonlinear dimensionality reduction technique frequently used to visualize high-dimensional data in two or three dimensions. The low dimensional representations are modeled such that nearby points, as defined by the Euclidean distance, are similar in the high dimensional space, while distant points are dissimilar with high probability, thus maintaining the relationship between surrounding points.

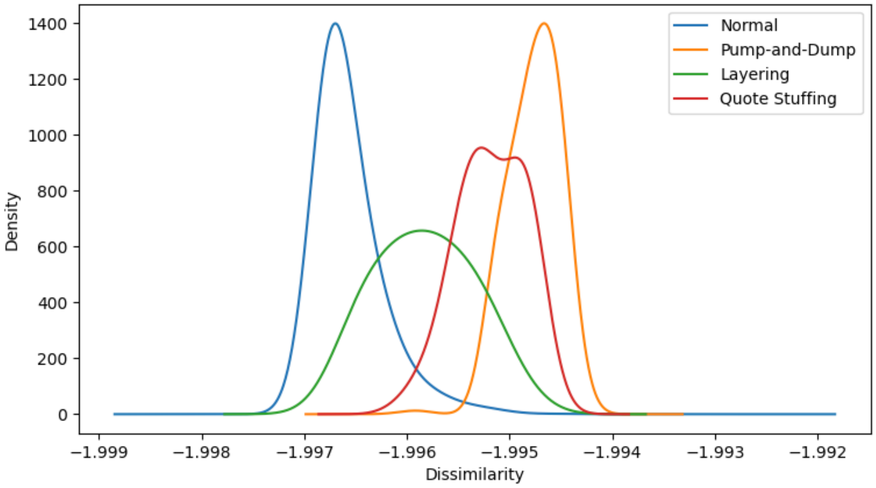

Fig. 25. Empirical distributions of our trained dissimilarity measure on the temporal representations for normal, pump-and-dump, layering, and quote stuffing subsequences.

Table 46. $p - values$ of two-sample Kolmogorov-Smirnov test for the first-order stochastic dominance of dissimilarity distributions found by the proposed hybrid model.

F(x) | G(x)	**Pump-and-Dump**	**Quote Stuffing**	**Layering**	**Normal**
Pump-and-Dump	-	1.000	1.000	1.000
Quote Stuffing	0.000	-	1.000	0.996
Layering	0.000	0.000	-	0.990
Normal	0.000	0.000	0.000	-

H_0: $F(x) \leq G(x)$, $\forall x$ and H_1: $F(x) > G(x)$ for at least one x, to evaluate the stochastic ordering between each pair of dissimilarity distributions. Firstly, we can conclude that pump-and-dump sequences have first-order stochastic dominance over all the order distributions, and quote stuffing dominates layering. Secondly, all trade–based manipulation dissimilarity distributions have first-order stochastic dominance over the normal dissimilarity distribution, further demonstrating that the hybrid model is indeed able to differentiate between normal and abnormal LOB sequences.

5.3. Ablation study

To better understand the importance of each aspect in the proposed methodology (Transformer vs. recurrent models, dissimilarity vs. estimation methods, temporal encoding vs. no encoding/non-sequential encoding), we contrast its performance against four similar model variants. The hybrid Transformer-AE/OC-SVM model (Dissimilarity Transformer) is compared to an equivalent LSTM-AE/OC-SVM model (Dissimilarity LSTM), and an equivalent MLP-AE/OC-SVM model (Dissimilarity MLP) that all have the same representation vector dimension $d = 128$.[37] The proposed framework is also evaluated against a simple OC-SVM (Dissimilarity only OC-SVM), which takes in the entire multivariate subsequence as a single vector, thereby ignoring the temporality of the data and any encoder dimensionality reduction. Finally, the dissimilarity approach is compared to the usual reconstruction method (Reconstruction Transformer), using the complete bottlenecked Transformer-AE trained for the hybrid approach, and ignoring the OC-SVM. Table 47 presents the performance of these methods on the test set containing all stocks, and all simulated frauds, for a general F_4-optimal τ^* per method.

Table 47. Overall performance of the proposed methodology on test set containing all stocks, compared to variant approaches. Best metric in bold, second best is underlined.

Method | Metric	**AUROC**	**AUPRC**	**F$_4$**	**Precision**	**Recall**
Dissimilarity Transformer	**0.900**	**0.847**	**0.935**	0.628	0.965
Dissimilarity LSTM	<u>0.877</u>	0.514	<u>0.869</u>	0.332	0.967
Dissimilarity MLP	0.722	0.628	0.767	0.162	1.000
Dissimilarity only OC-SVM	0.803	<u>0.652</u>	0.792	0.602	0.808
Reconstruction Transformer	0.467	0.431	0.782	0.174	1.000

As can be concluded from Table 47, the bottlenecked Transformer architecture is able to generate more descriptive representation vectors than LSTMs, facilitating the discriminative objective of the OC-SVM in the representation space, which in turn results in greater AUROC, AUPRC, and F_4-measure. We can also notice the importance of integrating the temporality of the data in its representation, as the MLP generates a worse performance than

[37] In this context, the term "equivalent" refers to a similar L_2 reconstruction loss obtained by the autoencoders on the validation set.

the Transformer encoder, even lower than the OC-SVM alone. This demonstrates that simple encoding is not enough, and that the performance of the bottleneck-Transformer model results mainly from its ability to learn rich temporal representations. Finally, the estimation method using the reconstruction error of the same bottlenecked Transformer autoencoder results in poor detection accuracy compared to all dissimilarity methods. Overall, Table 47 demonstrates that, out of multiple similar variants, the combination of temporal encoding from a Transformer model and a dissimilarity approach generates the best performance in terms of AUROC, AUPRC, and F_4-measure, in the context of LOB time series anomaly detection.

6. Conclusion

In this article, we propose a novel time series anomaly detection model tailored to LOB time series, with an application to trade–based manipulation detection in five NASDAQ stocks. We introduce a new autoencoder, the bottlenecked Transformer autoencoder, which can learn semantically rich temporal representations of LOB time series. The representation space of its encoder eases the separability of normal and abnormal LOB behavior, allowing a one-class classification algorithm to discriminate between the two categories with a dissimilarity function, thus detecting accurately anomalous LOB subsequences out-of-sample. The model utilizes a greater pool of LOB features to capture a larger range of fraud types compared to the previous literature by integrating the price, volume, and time dynamics of the LOB, a necessary step in the evolution of trade–based manipulation detection (Khodabandehlou and Golpayegani [26]). Finally, the framework achieves new state-of-the-art performance by adapting recent deep learning methods proposed for image anomaly detection, reducing the gap between this active field and the financial market anomaly detection literature.

We also present a complete trade–based manipulation scenario simulator able to generate pump-and-dump, layering, and quote stuffing tactics. The random scenarios are used to quantify the performance of the anomaly detection model. This is an important departure from earlier literature in financial fraud detection, since it relied on repeating the same limited set of orders, hence overestimating the performance of previous methods. We show that the proposed deep unsupervised anomaly detection model captures these three types of fraud on all five stocks, meaning that it learns an asset–independent notion of normal LOB behavior, without needing any prior knowledge of fraudulent patterns. We also empirically show that the Transformer–based model learns better representations than the popular LSTM network, and that dissimilarity methods outperform more traditional estimation–based anomaly detection on LOB time series. Furthermore, we quantify the difficulty of detecting pump-and-dump, layering, and quote stuffing manipulations, a first in the literature. Providing this kind of analysis is also helpful in determining the comparative strengths and weaknesses of new trade–based manipulation detectors, in addition to traditional performance metrics.

The proposed framework is a strong alternative to the rule–based systems currently used by market regulators (Golmohammadi and Zaine [20]) in two ways. First, it learns a general notion of normalcy, so a single model instance can be utilized for any asset, and to detect any type of anomalous behavior. Second, it can dynamically adapt to market regimes, whereas rule–based systems need to be manually adjusted. But, market data drift is an important aspect to consider when deploying any data–based model (Žliobaitė et al. [58]), because the past learned notion of normality might slowly depart from future normal market behaviors. It is primordial to know when to retrain anomaly detection frameworks, and further research in that sense, in the context of financial markets, is important.

Also, semi-supervised learning techniques for anomaly detection have recently been proposed in the deep learning literature (e.g., Ruff et al. [48]), where small sets of known anomalies are used to train the models in conjunction with the unlabeled data, boosting their detection performance over pure unsupervised methods. It would be worthwhile to explore and adapt these methods to the trade–based manipulation detection problem, as only limited collections of frauds are available to researchers. Additionally, semi-supervised approaches open the door to human-in-the-loop models where they could learn from an ever-growing pool of detected frauds confirmed by market regulators, constantly raising their detection capabilities. Our framework can act as a starting point on which the semi-supervised methods can be built upon.

Acknowledgments

Financial support from Fin-ML, the Montreal Exchange (TMX), Centre d'intelligence en surveillance des marchés financiers, and Mitacs is graciously acknowledged. We also thank the TMX regulatory division team and IVADO for their constant support in this project.

References

[1] Abbas, B., Belatreche, A., and Bouridane, A. (2019). Stock price manipulation detection using empirical mode decomposition based kernel density estimation clustering method. *Intelligent Systems and Applications: Proceedings of the 2018 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 851–866.

[2] Aggarwal, R. and Wu, G. (2006). Stock market manipulation. *The Journal of Business*, 79(4):1915–1953.

[3] Agrawal, S. and Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, 60:708–713.

[4] Audibert, J., Michiardi, P., Guyard, F., Marti, S., and Zuluaga, M. A. (2020). USAD: Unsupervised anomaly detection on multivariate time series. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3395–3404.

[5] Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, 54(3):1–33.

[6] Cao, Y., Li, Y., Coleman, S., Belatreche, A., and McGinnity, T. (2013). A hidden Markov model with abnormal states for detecting stock price manipulation. *IEEE Transactions on Neural Networks and Learning Systems*, 26(2):318–3019.

[7] Cao, Y., Li, Y., Coleman, S., Belatreche, A., and McGinnity, T. (2014). Detecting price manipulation in the financial market. In *IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr)*, pages 77–84. IEEE.

[8] Chalapathy, R. and Chawla, S. (2019). Deep learning for anomaly detection. Available at https://arxiv.org/abs/1901.03407.

[9] Choi, K., Hawthrone, C., Simon, I., Dinculescu, M., and Engel, J. (2020). Encoding musical style with transformer autoencoders. *International Conference on Machine Learning*, pages 1899–1908.

[10] Chullamonthon, P. and Tangamchit, P. (2022). A transformer model for stock price manipulation detection in the stock exchange of Thailand. *IEEE International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).*

[11] Chullamonthon, P. and Tangamchit, P. (2023). Ensemble of supervised and unsupervised deep neural networks for stock price manipulation detection. *Expert Systems with Applications*, 220:119698.

[12] Close, L. and Kashef, R. (2020). Combining artificial immune system and clustering analysis: A stock market anomaly detection model. *Journal of Intelligent Learning Systems and Applications*, 12:83–108.

[13] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[14] Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27.

[15] Deng, A. and Hooi, B. (2021). Graph neural network–based anomaly detection in multivariate time series. *Proceedings of the AAAI conference on artificial intelligence*, 35(5):4027–4035.

[16] Diaz, D., Theodoulidis, B., and Sampaio, P. (2011). Analysis of stock market manipulations using knowledge discovery techniques applied to intraday trade prices. *Expert Systems with Applications*, 38(10):12757–12771.

[17] Egginton, J., Ness, B. V., and Ness, R. V. (2016). Quote stuffing. *Financial Management*, 45(3):583–608.

[18] Forney, G. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278.

[19] Gandhmal, D. and Kumar, K. (2019). Systematic analysis and review of stock market prediction techniques. *Computer Science Review*, 34.

[20] Golmohammadi, K. and Zaine, O. (2015). Time series contextual anomaly detection for detecting market manipulation in stock market. *International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

[21] Golmohammadi, K., Zaine, O., and Diaz, D. (2014). Detecting stock market manipulation using supervised learning algorithms. *International Conference on Data Science and Advanced Analytics (DSAA)*, pages 435–441.

[22] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[23] Huang, R. and Polak, T. (2011). Lobster: Limit order book reconstruction system. Available at https://ssrn.com/abstract=1977207.

[24] IIROC (2013). IIROC Guidance on Certain Manipulative and Deceptive Trading Practices. Available at https://www.iiroc.ca/news-and-publications/notices-and-guidance/guidance-certain-manipulative-and-deceptive-trading-practices. Accessed March 13, 2023.

[25] Kamps, J. and Kleinberg, B. (2018). To the moon: Defining and detecting cryptocurrency pump-and-dumps. *Crime Science*, 7(18).

[26] Khodabandehlou, S. and Golpayegani, S. (2022). Market manipulation detection: A systematic literature review. *Expert Systems with Applications*, page 118330.

[27] Kingma, D. P., J., and Ba (2014). Adam: A method for stochastic optimization. Available at https://arxiv.org/abs/1412.6980.

[28] Leangarun, T., Tangamchit, P., and Thajchayapong, S. (2016). Stock price manipulation detection based on mathematical models. *International Journal of Trade, Economics and Finance*, pages 81–88.

[29] Leangarun, T., Tangamchit, P., and Thajchayapong, S. (2018). Stock price manipulation detection using generative adversarial networks. *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2104–2111.

[30] Leangarun, T., Tangamchit, P., and Thajchayapong, S. (2021). Stock price manipulation detection using deep unsupervised learning: The case of Thailand. *IEEE Access*, 9:106824–106838.

[31] Lee, E., Eom, K., and Park, K. (2013). Microstructure-based manipulation: Strategic behavior and performance of spoofing traders. *Journal of Financial Markets*, 16:227–252.

[32] Li, D., Chen, D., Jin, B., Shi, L., Goh, J., and Ng, S. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series, 28th International Conference on Artificial Neural Networks*, pages 703–716.

[33] Lin, T. (2016). The new market manipulation. *Emory Law Journal*, 66:1253.

[34] Maaten, L. V. D. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11).

[35] Meng, H., Zhang, Y., Li, Y., and Zhao, H. (2020). Spacecraft anomaly detection via transformer reconstruction error. In *Proceedings of the International Conference on Aerospace System Science and Engineering 2019*, pages 351–362. Springer.

[36] Montero, I., Pappas, N., and Smith, N. (2021). Sentence bottleneck autoencoders from transformer language models. Available at https://arxiv.org/abs/2109.00055.

[37] Nanex Research (2010). Latency on Demand? Available at http://www.nanex.net/FlashCrash/FlashCrashAnalysis_LOD.html. Accessed March 15, 2023.

[38] Nanex Research (2011). The WAB Event. Available at http://www.nanex.net/StrangeDays/12142011.html. Accessed March 15 2023.

[39] Nanex Research (2014). The Quote Stuffing Trading Strategy. Available at http://www.nanex.net/aqck2/4670.html. Accessed March 15, 2023.

[40] Pang, G., Shen, C., Cao, L., and Hengel, A. (2021). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38.

[41] Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[43] Reiss, T., Cohen, N., Bergman, L., and Hoshen, Y. (2021). PANDA: Adapting pretrained features for anomaly detection and segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2806–2814.

[44] Reiss, T., Cohen, N., Horwitz, E., Abutbul, R., and Hoshen, Y. (2023). Anomaly detection requires better representations. *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 56–68.

[45] Rizvi, B., Belatreche, A., and Bouridane, A. (2019). A dendritic cell immune system inspired approach for stock market manipulation detection. *IEEE Congress on Evolutionary Computation (CEC)*, pages 3325–3332.

[46] Rizvi, B., Belatreche, A., Bouridane, A., and Mistry, K. (2020a). Stock price manipulation detection based on autoencoder learning of stock trades affinity. *IEEE International Joint Conference on Neural Networks (IJCNN).*

[47] Rizvi, B., Belatreche, A., Bouridnae, A., and Watson, I. (2020b). Detection of stock price manipulation using kernel based principal component analysis and multivariate density estimation. *IEEE Access 8*, pages 135989–136003.

[48] Ruff, L., Vandemeulen, R., Görnitz, N., Binder, A., Müller, E., Müller, K., and Kloft, M. (2020). Deep semi-supervised anomaly detection. *Eight International Conference on Learning Representations (ICLR).*

[49] Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In *International Conference on Machine Learning (ICML)*, pages 4393–4402. PMLR.

[50] Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS one*, 10(3):e0118432.

[51] Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., and Platt, J. (1999). Support vector method for novelty detection. *Advances in Neural Information Processing Systems*, 12:582–588.

[52] Shen, L., Li, Z., and Kwok, J. (2020). Time series anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33:13016–13026.

[53] Siering, M., Clapham, B., Engel, O., and Gomber, P. (2017). A taxonomy of financial market manipulations: Establishing trust and market integrity in the financialized economy through automated fraud detection. *Journal of Information Technology*, 32:251–269.

[54] Sohn, K., Li, C., Yoon, J., Jin, M., and Pfister, T. (2021). Learning and evaluating representations for deep one-class classification. *International Conference on Learning Representations (ICLR).*

[55] Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., and Pei, D. (2019). Robust anomaly detection for multivariate time series through stochastic recurrent neural network. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2828–2837.

[56] Tax, D. and Duin, R. (2004). Support vector data description. *Machine Learning*, 54:45–66.

[57] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[58] Žliobaitė, I., Pechenizkiy, M., and Gama, J. (2016). An overview of concept drift applications. In N. Japkowicz, N. and Stefanowski, J., editors, *Big Data Analysis: New Algorithms for a New Society*, pages 91–114. Springer International Publishing, Cham.

[59] Wang, Q., Wu, W., Huang, X., and Yang, K. (2019). Enhancing intraday stock price manipulation detection by leveraging recurrent neural networks with ensemble learning. *Neurocomputing*, 347:46–58.

[60] Wang, X., Zhao, Y., and Pourpanah, F. (2020). Recent advances in deep learning. *International Journal of Machine Learning and Cybernetics*, 11:747–750.

[61] Xu, J., Wu, H., Wang, J., and Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. *International Conference on Learning Representations (ICLR)*.

[62] Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., and Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:1409–1416.

[63] Zhao, H., Wang, Y., Duan, J., Huang, C., Cao, D., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q. (2020). Multivariate time-series anomaly detection via graph attention network. *IEEE International Conference on Data Mining (ICDM)*, pages 841–850.

[64] Ögüt, H., Doğanay, M., and Aktaş, R. (2009). Detecting stock price manipulation in an emerging market: The case of Turkey. *Expert Systems with Applications*, 36(9):11944–11949.

Conclusion

This thesis proposes novel methods able to exploit the vast informational content of limit order book time series in the context of statistical arbitrage, and anomaly detection.

The first two chapters contribute to the better understanding of high-frequency traders' arbitrage activities. Chapter 1 explores statistical and mechanical arbitrage mechanisms in cross-listed assets. Its main contribution lies in the construction of synthetic instruments possessing desirable statistical properties, namely, stationarity and mean-reversion, from which theoretical arbitrage bounds can be inferred for cross-listed assets. An original trading strategy captures price deviations outside these bounds and is empirically demonstrated to be profitable in North American markets in 2019. Moreover, this chapter introduces a practical backtesting methodology enabling the study of information latency and its effect on the profitability of high-frequency trading. This framework is applied to previous strategies developed for cross-listed assets, showing that ignoring market frictions results in severe losses and demonstrating their importance for future academic research. Chapter 1 also discusses how to generalize the arbitrage bounds and strategy to more trading environments. It would be worthwhile to explore the proposed methodology in multiple-listed assets, and more than two currencies, as to have a more global understanding of international arbitrage.

Chapter 2 investigates high-frequency lead-lag relationships, and its contribution is twofold. First, it introduces a novel econometric model specifically designed to exploit the lead-lag effect between two related assets that is more accurate out-of-sample compared to previous ones. Second, a new high-frequency trading strategy is devised to capture the arbitrage opportunities detected by the econometric model in European markets in 2013. The overall contribution of this chapter is in the empirical demonstration that lead-lag relationships are a source of viable arbitrage, even when the market frictions studied in Chapter 1 are introduced. The existence of lead-lag relationships has been shown in most financial assets, but their economic importance has been questioned for almost a decade. This chapter is the first to formally, and pragmatically, show the potential profitability behind them, thus unveiling market inefficiencies exploitable by high-frequency traders. The predictive model focuses on the price dynamics between the leading and lagging assets. It would be interesting to also integrate their timing dynamics, which could lead to more accurate return predictions and better risk management for the trading strategy. Another fascinating research avenue would be to explore lead-lag relationships in the context of market making, where they could probably be useful for market makers who might better anticipate price movements on lagging assets.

Finally, Chapter 3 focuses on the applicability of modern deep learning and machine learning methods in algorithmic trade–based manipulation detection. It contributes to the financial literature by proposing a novel Transformed–based unsupervised anomaly detection framework able to learn an asset–independent notion of normalcy, thus capturing any type of fraud, and more generally, any anomaly, without requiring prior knowledge of fraudulent patterns or labeled data. Furthermore, the framework integrates a more general set of limit order book features, hence expanding the types of frauds that can be captured simultaneously by a single model. A more exhaustive trade–based manipulation simulation approach is also proposed, rendering the detection results more reliable compared to past studies. It is shown that the framework achieves the best detection results when compared to previous methods on simulated fraud scenarios in NASDAQ stocks. Semi-supervised learning, where a small set of labels is given to a model in conjunction with unlabeled data to increase its detection capabilities, would be a valuable next step in financial market manipulation detection. It is possible to adapt the proposed methodology to the semi-supervised paradigm, which then opens the door to human-in-the-loop models for market regulators.