Statistical methods in financial market dynamics and portfolio strategies.

Qi Jin
St. Anne's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
July 2025

1 | Acknowledgment

I would like to express my deep gratitude to my supervisors, Álvaro Cartea and Mihai Cucuringu for their support and encouragement throughout the past 30 months. Their guidance, patience, and unwavering belief in my work have been instrumental throughout my DPhil journey. I am especially grateful for the freedom they gave me to explore my ideas, and for their genuine support and enthusiasm for the topics I chose to pursue.

I am also grateful for my excellent research collaborators. Working with Professor Mungo Wilson has been a privilege — his familiarity with the literature and constructive feedback pushed our work to a higher standard. I am equally thankful to Ruotong Cao, Yuantao Shi, Jiexiu Zhu, whose rigor, energy, and openness made our collaboration productive and enjoyable. It is especially meaningful to see our joint work contribute to their theses in the future, and I look forward to seeing their researches continue to grow. I would also like to thank the examiners of my viva Liying Wang and Shumiao Ouyang for their profound expertise, meticulous examination, and helpful comments on my thesis.

I am indebted to the Oxford-Man Institute of Quantitative Finance (OMI), Man Group, and St. Anne's College for funding my studies and for all other administrative support. I am also extremely thankful to my parents for their unconditional support and love throughout my past eleven years of studying abroad.

I would also like to extend my gratitude to all my colleagues at the OMI for inspiring discussions and emotional support. I am especially thankful to Yutong Lu, Xingyue Pu, and Chao Zhang, who were seniors when I first joined. Their guidance and friendship provided invaluable support during my earlier years.

I am thankful to my peers Chuqiao Lin, Qiucheng Qian, Shuke Wang, Yunyi Wang, and Haoyang Zhang along with others for being amazing friends and for many moments of laughter and encouragement that made this journey more meaningful and enjoyable.

Finally, I would like to acknowledge the artists and writers whose works have been a constant source of inspiration and solace. The writings of Albert Camus, Hermann Hesse, and Virginia Woolf, the music of Junjie Lin, and the films of Wong Kar Wai have accompanied me through emotional highs and lows ever since I first arrived at Oxford as a teenager, six and a half years ago. Their art has shaped the way I see the world and sustained me in more ways than I can express.

2 | Abstract

This thesis uses statistical methods to explore topics in financial economics. In particular, we focus on topics related to financial market dynamics and portfolio strategies. This thesis makes four contributions to the literature.

First, we introduce a method to detect linear and nonlinear lead-lag relationships in stock returns that uses pairwise Lévy-area and cross-correlation to rank assets from leaders to followers. We construct portfolios by trading followers based on leaders' prior returns, hedged with an SPY ETF. With data from 1963 to 2022 for over 500 stocks, our portfolios achieve annualized returns over 20% and Sharpe ratios over 2. The relationships we discover are only partially explained by traditional factors like size or sector. Our results support the slow information diffusion hypothesis as daily rebalanced portfolios outperform less frequently rebalanced ones.

Second, we study the effect of intraday volume shocks on stock returns during overnight and intraday periods. We discover a significant positive relationship between volume shocks and subsequent overnight returns, while no such effect exists during the next intraday session. Well-known asset pricing risk factors and common explanations that associate abnormal trading volume with investor attention and cost of capital cannot account for the distinct intraday and overnight patterns we observe. We employ linear and machine learning models to forecast volume shocks and to construct portfolios that monetize the positive correlation between volume shocks and overnight stock returns. Our approach addresses the issue that volume shock is only known after the close auction; we show that this issue of non-tradability does not explain the observed relationship between volume shock and overnight stock returns.

Third, we propose a framework to construct statistical arbitrage portfolios with graph clustering algorithms. First, we use five clustering methods to partition the correlation matrix of market residual returns of stocks into clusters. Next, we construct and evaluate the performance of mean-reverting statistical arbitrage portfolios within each cluster. We show that our proposed framework generates profitable trading strategies with over 10% annualized returns and statistically significant Sharpe ratios above one. The performance of our statistical arbitrage portfolios is neutral to the market and cannot be fully explained by intra-industry mean-reversion effects.

In the last part, we examine the investment value in sell-side analyst price targets. We treat each analyst as a portfolio manager and use their price targets to construct 12-month implied return forecasts and self-financing long-short portfolios for each analyst. Our empirical analysis shows that while the average analyst does not generate statistically significant alpha relative to the returns of a long-only portfolio benchmark, a subset of analysts exhibits persistent alpha. Motivated by this heterogeneity, we introduce a "fund-of-analysts" framework that first predicts analyst performance and then dynamically allocates weights across analysts based on predicted analyst performances. Our results show that this meta-portfolio strategy can yield significant alpha over long-only benchmarks.

Table of Contents

List of Figures

List of Tables

3 | Introduction

The quest to identify systematic patterns in financial asset returns remains a central focus in financial economics (Feng et al., 2020; Kakushadze, 2016; Lo and MacKinlay, 1990). While the efficient market hypothesis posits that prices reflect all available information, an extensive literature documents persistent return predictability arising from behavioral biases, information frictions, and institutional constraints (Frazzini and Cohen, 2008; Lo and MacKinlay, 1990; Badrinath et al., 1995). In response, researchers have developed increasingly sophisticated methods—ranging from classical time-series econometrics to advanced machine learning—to study asset pricing dynamics and design profitable trading strategies (Zhang et al., 2020; Gu et al., 2019; He and Krishnamurthy, 2013; Kelly et al., 2022).

This thesis contributes to this broad literature by presenting four empirical studies that investigate informational frictions and return predictability in equity markets. Each project explores a different facet of return dynamics through data-driven statistical techniques that blend theory, statistical methods, and algorithmic tools. The overarching aim is to understand how information arrives, propagates, and ultimately influences the cross-section of stock returns, and to propose trading strategies that exploit inefficiencies in that process. In particular, the thesis focuses on four problems: detection of lead-lag relationships in stock returns (Chapter 4), understanding the relationship between intra-day trading volume and overnight stock returns (Chapter 5), neutralize portfolio exposure to latent factors in statistical arbitrage portfolios (Chapter 6), and understanding if there is investment value in sell-side analyst reports on individual stocks (Chapter 7)

A prominent strand of the stock return prediction literature posits that asset-price behavior

often exhibits cross-autocorrelations or lead-lag effects, driven by gradual information diffusion, investor inattention, and institutional trading activity (Chordia and Swaminathan, 2000; Lo and MacKinlay, 1990; Hou, 2007). Traditional approaches to uncover such relationships often invoke linear dependency assumptions or firm-level characteristics such as market capitalization and liquidity (Badrinath et al., 1995; Brennan et al., 1993; Amihud, 2002). In contrast, my first project proposes a methodology that uses the Lévy-area and cross-correlation of daily stock returns—without imposing *a priori* linear functional forms—to identify and rank pairs of stocks exhibiting significant lead-lag interactions. Our empirical findings suggest that a long-short strategy built around these identified pairs achieves annualized returns exceeding 20%. This result offers direct support for the slow-information-diffusion hypothesis and the notion that certain stocks systematically anticipate others (Frazzini and Cohen, 2008; Hou, 2007).

The second project extends the literature on trading activity and its impact on asset returns (Chordia et al., 2001; Barber and Odean, 2008; Amihud, 2002). Specifically, we examine how large, unexpected intra-day trading volume predicts overnight stock returns. By analyzing trading volume data and overnight stock returns, we document that stocks with large volume surprise during trading hours tend to have higher returns overnight when the market is closed, but these effects dissipate once the market opens on the next trading day. To make these patterns that we document tradable, we develop a machine learning framework that forecasts volume shocks, leading to a suite of simple yet profitable trading strategies on stock overnight returns. Our findings highlight the importance of volume dynamics in understanding overnight return behavior and inform how market participants might better position themselves each day before market close.

The third project addresses a perennial challenge in finance: constructing market-neutral,

mean-reverting portfolios to pursue statistical arbitrage opportunities (Avellaneda and Lee, 2010; Elliott et al., 2005). We propose a clustering-based approach that builds on graph clustering algorithms designed to handle nuanced correlation structures, including those found in large cross-sections of equities (Ng et al., 2001; Cucuringu et al., 2019b,a). By grouping stocks into clusters that share latent risk factors or display similar co-movement patterns, we design self-financing, mean-reverting trading strategies within each cluster. Our results indicate that such portfolios can generate robust annualized returns exceeding 10% with Sharpe ratios above one. These portfolios exhibit minimal directional exposure, underscoring the viability of this type of clustering for risk management and diversification.

The final project investigates the informational content of sell-side analyst price targets by treating analysts as portfolio managers whose forecasts imply actionable signals. While prior work largely focus on firm-level implications of analyst research (Womack, 1996; Clement, 1999; Brav and Lehavy, 2003; Asquith et al., 2005; Bradshaw et al., 2012), we adopt an analyst-centric perspective and construct long-short portfolios based on each analyst's target-price-implied returns. We evaluate analyst performance against passive benchmarks and uncover substantial heterogeneity in forecasting skill. Building on this, we propose a "fund-of-analysts" framework that first uses characteristics such as past performance and cross-sectional rank of analysts to predict future analyst returns and then dynamically allocates capital across analysts accordingly. Our findings contribute to the literature on analyst skill, forecast-based investing (Stickel, 1992; Tamura, 2002; Gleason and Lee, 2003; Dechow and You, 2020; Loudis, 2024), and the broader understanding of how investors can extract value from heterogeneous sell-side research.

Collectively, the four projects forge a connection between algorithmic innovation, and asset-pricing practice. They reflect a broader shift in asset pricing research: away from

structural models with rigid assumptions and toward empirical models that are flexible, data-driven, and robust to model uncertainty (Gu et al., 2019; Feng et al., 2020). By employing techniques such as lead-lag relationship detection methods, machine learning, clustering, and forecast aggregation, this thesis sheds light on persistent anomalies in return behavior and proposes scalable strategies to exploit them.

The remainder of the thesis is organized as follows: Chapter 4 discusses lead-lag relationships in stock returns; Chapter 5 details the the relationship between overnight stock returns and intraday volume shocks; Chapter 6 focuses on correlation matrix clustering for statistical arbitrage portfolios; Chapter 7 studies the investment value in analyst price targets. The appendix contains additional information on robustness, data exploration, and alternative specifications of parameters.

4 | Detecting Lead-Lag Relationships in Stock Returns and Portfolio Strategies

4.1 Introduction

Changes in stock prices of some firms tend to follow those of other firms. This relationship between stock prices is often referred to as a *lead-lag* relationship. Detecting lead-lag relationships among a large set of stocks is not straightforward. The extant literature uses ad-hoc methods to select leaders and followers, and employs these two sets of stocks in investment strategies to evaluate the economic significance of the lead-lag relationship. For example, Lo and MacKinlay (1990) assume that large market capitalization stocks lead small market capitalization stocks. They build equal-weighted portfolios within each quantile of market capitalizations and use the cross-autocorrelation between the five portfolios to evaluate the trading performance of the lead-lag relationship. Empirical evidence suggests that firm size (Lo and MacKinlay (1990)), trading volume (Chordia and Swaminathan (2000)), institutional ownership (Badrinath et al. (1995)), and other firm characteristics contribute to the lead-lag identity of a stock.

Empirically, however, many lead-lag relationships change over time and often cannot be explained by sorting stocks on a single firm characteristic.[1] Therefore, it is necessary to detect, instead of assume and then verify, lead-lag relationships.

Our objective is to find lead-lag relationships without explicitly assuming a link between firm characteristics and lead-lag relationships; specifically, we identify leaders and followers

[1] See Appendix A.1 for motivating examples.

based on the relationship in stock returns over the past days, and we assume that the discovered lead-lag relationship continues to exist in the near future. To do this, we develop a data-driven method that employs stock returns to identify leaders and followers, and we show that the lead-lag relationships we find are economically significant. We achieve this in three steps. First, we design an algorithm that identifies the direction and strength of the lead-lag relationship between the returns of two stocks. Second, we propose a framework that uses state-of-the-art algorithms to rank stocks from leaders to followers based on the pairwise relationships. Third, we construct a zero-cost portfolio to assess the returns predictability of the leaders over the followers, and we measure the economic significance of the portfolio's performance.

Specifically, in the first step we design a method to score the lead-lag relationship between pairs of assets. The sign of the score indicates which of the two assets is more likely the leader, and the magnitude of the score quantifies the strength of the lead-lag relationship. We compute the pairwise score for all combinations of asset pairs, and use the pairwise scores to construct a skew-symmetric matrix, which we refer to as the lead-lag matrix. In this paper, we introduce three scoring methods. Two methods use cross-correlations to capture linear lead-lag relationships; and the third is based on the Lévy-area of pairwise asset returns, which captures both linear and nonlinear lead-lag relationships.

In the second step, we use the lead-lag matrix to sort the stocks from "most likely to lead" to "most likely to follow". First, we compute the mean of each column of the skew-symmetric lead-lag matrix and order the columns according to their mean, from highest to lowest. A priori, stocks in columns with high means are more likely to be leaders, and those in columns with low means are more likely to be followers.

Finally, after identifying a set of leaders and a set of followers, we use the sign of the previous return of the set of leaders as a signal to buy or to sell an equal-weighted portfolio of the followers. The portfolio is financed by taking an offsetting position on the SPY ETF, so the value of the portfolio at inception is zero. Every day, we rank the stocks to rebalance the portfolio, and we compare the portfolio's performance with those of the lead-lag-based portfolios proposed in Lo and MacKinlay (1990), Chordia and Swaminathan (2000), and Hou (2007). For robustness and comparisons, we also study portfolios that are rebalanced at bidiurnal, weekly, bi-weekly, tri-weekly, and monthly frequencies — stocks are ranked at the same frequency as that of the portfolio rebalances.

In line with the extant literature, our study uses the profitability of the lead-lag portfolio to evaluate the economic significance of the lead-lag relationship between pairs and between groups of stocks. As an evaluation metric, the profitability of the portfolio has two advantages. One, portfolio profitability provides a direct comparison with the existing literature. Two, as suggested by Kelly et al. (2022), directly considering portfolio performance is more suitable than using the R^2 of a linear regression. A portfolio can generate significant economic profits even if the predicted R^2 of a regression on returns is negative because the variance of forecasts can heavily influence the predicted R^2.

We use data from 1963 to 2022 to compare the performance of our lead-lag portfolios with those of the portfolios built with the lead-lag relationships identified in the literature. We use only the stocks with large market capitalization to ensure that our portfolios are tradable. Specifically, in each trading day, our analysis includes approximately 550 stocks which are the top 25% largest stocks by market capitalization in NYSE, NASDAQ, and AMEX. Our results show that the portfolios based on the lead-lag relationships we detect outperform those proposed in the extant literature that are based on market capitalization,

trading volume, or sector identity. Specifically, the annualized returns of the portfolios we form with the leaders and followers we detect are over 20% and the Sharpe ratio of the portfolios are above 2. The annualized return and Sharpe ratio of our portfolios are, respectively, three times and approximately twice those of the portfolios proposed in the extant literature. The performance of our portfolios is robust to various choices of parameter values, ranking algorithms, and selection of sets of tradable stocks. In particular, we find that the identity of a stock as a leader or follower changes frequently over time. A lead-lag relationship found at any particular time window rarely remains profitable after 6 months because the lead-lag relationship changes. Our results are economically significant after accounting for industry standard transaction costs and trading frictions. The return of our portfolios cannot be explained by the Fama–French risk factors, illiquidity of stocks, or lead-lag relationships proposed in the literature. In particular, there is little overlap between the leaders and the followers that we identify and those identified by the literature; thus, our method detects fundamentally different lead-lag relationships from those identified in the literature.

In addition, we study one source of lead-lag relationships in stock prices. The literature suggests that asynchronous reactions in stock prices to common information cause the lead-lag relationships; they refer to this conjecture as the hypothesis of slow information diffusion.[2] The sluggishness in reaction to common information depends on various features, including relatively small firm size (Lo and MacKinlay (1990)), relatively low trading volume (Chordia and Swaminathan (2000)), relatively low institutional ownership (Badrinath et al. (1995)), and many other types of market features.[3]

[2]Lo and MacKinlay (1990), Brennan et al. (1993), and Badrinath et al. (1995) argue that lead-lag effects exist because some firms react more slowly to common information than others; Hou (2007) is the earliest paper we found that explicitly calls this the “slow information diffusion hypothesis".

[3]The literature proposes the “limited investor attention" hypothesis to explain why some stocks take longer than others to react to common information.

Our results lend strong support to the slow information diffusion hypothesis. We perform a temporal analysis of our lead-lag portfolios to study the speed of information diffusion. If the slow information diffusion hypothesis is valid, then the economic significance of the lead-lag relationships we identify would gradually diminish as the time gap between identifying the lead-lag relationship and executing the trades to build the portfolio increases. In our study, we vary the rebalancing frequency of the lead-lag portfolios and the frequency of detecting lead-lag relationships from daily to bidiurnal, weekly, bi-weekly, tri-weekly, and monthly. Our results show that as the rebalancing frequency decreases, the performance of the lead-lag portfolios decreases. In the 1960s and early 1970s, the performance of the portfolios with daily rebalancing is indistinguishable from that of the portfolios rebalanced at slower frequencies. However, over time, the daily portfolio gradually outperforms the monthly portfolio, and then the tri-weekly, the bi-weekly, and the weekly rebalanced portfolios from the late 1970s onward.

Our findings contribute to the literature on return predictability, specifically the strand that explores lead-lag relationships among stocks. After Lo and MacKinlay (1990), there has been a growing literature on the lead-lag relationships between two portfolios of stocks. In contrast to our approach, the extant literature does not develop methods to detect which stocks are leaders and which are followers; instead, it assumes which stocks lead and which follow, and uses portfolios to evaluate the lead-lag relationships. Badrinath et al. (1995) study the lead-lag relationship between groups of stocks with higher institutional ownership (i.e., leaders) and lower institutional ownership (i.e., followers), while Chordia and Swaminathan (2000) study the lead-lag relationship where groups of stocks with larger trading volume lead stocks with smaller trading volume. Brennan et al. (1993) verify that stocks with higher investment analyst coverage lead those with lower coverage. Hou

(2007) shows lead-lag relationships between large and small market capitalization stocks within each industry, while Frazzini and Cohen (2008) and Menzly and Ozbas (2009) assume lead-lag relationships between groups of stocks with economic links specified by consumer-supplier relationships from financial reports data. More recently, Parsons et al. (2020) assume lead-lag relationships between groups of stocks from different sectors that are co-headquartered, and Huang et al. (2022) show that stocks with frequent, gradual price updates tend to lead those with infrequent, sharp price updates.

We also contribute to a more general literature that uses the cross-section of stock returns to predict asset price movements. For example, DeMiguel et al. (2014) build a vector autoregression (VAR) model that regresses asset returns on the cross-section of previous returns of all other stocks and build arbitrage portfolios. Kelly et al. (2023) provide a theoretical framework on building optimal portfolios based on the cross-section of stock returns, and Yan and Yu (2023) study cross-stock momentum portfolios based on the framework. We contribute to this literature by introducing a framework to rank leadership in stocks and use the returns of the high-ranked stocks to predict the returns of the low-ranked stocks.

In contrast to the strand of finance literature on lead-lag relationships, the literature in network science and graph-based machine learning develops data-driven methods to detect lead-lag relationships and applies them to networks of financial assets. For example, Bennett et al. (2022) use clustering algorithms to classify stocks into several groups and study lead-lag relationships between the stock groups. Li et al. (2021) use the number of days for which the magnitude of returns of one stock is similar to that of the previous return of another stock to construct a network and fit power-law distributions to determine leaders

and followers.[4] This strand of the literature uses financial data to test their proposed methods, but does not study the economic principles of lead-lag relationships, see also Shi et al. (2023) and Zhang et al. (2023).

To the best of our knowledge, our proposed methodology is the first data-driven method in the finance literature designed to detect lead-lag relationships and to verify the slow information diffusion hypothesis in stock returns on daily or slower frequencies. However, several other lines of work explore lead-lag relationship detection for time series in fields such as statistics, econometrics, and machine learning. We fill the gap in this literature by designing a data-driven method customized to financial time series. A well-known parametric method in econometrics to detect lead-lag relationships is the Granger Causality test, Granger (1969). Scherbina and Schlusche (2018) use this method to test return lead-lag pairs among individual stocks, and prove pairwise return predictability upon news release, while Basnarkov et al. (2020) use this method to test for lead-lag relationships in foreign exchange markets. Similarly, although not designed for lead-lag relationships, the Sargan–Hansen test can be used to test lead-lag relationships when the link is considered as an instrumental causal relationship, see Sargan (1958). This strand of the literature assumes a linear relationship between the leaders and the followers.

While studies in the lead-lag literature assume linear relationships between leaders and followers, studies in other strands of finance suggest that nonlinear relationships should be considered. Works such as Campbell and Cochrane (1995); Bansal and Yaron (2004); He and Krishnamurthy (2013); DeMiguel et al. (2023), uncover nonlinear relationships in asset returns across various markets, frequencies, and asset classes; in particular, Pohl et al. (2018) suggest that linear approximations to nonlinear models can lead to consider-

[4]Some other works use methods in network science to build graphs with intra-day data and inspect lead-lag relationships, see Curme et al. (2015), Basnarkov et al. (2020), and Buccheri et al. (2019).

able errors in the model predictions of asset returns. DeMiguel et al. (2023) show that by accounting for nonlinearity, investors can construct portfolios of funds that incur positive alpha, whereas linear methods incur near-zero alpha. Altogether, these works illustrate the need to account for nonlinearity in lead-lag relationships. Our approach is fundamentally different from previous methods in econometrics because we do not assume a linear relationship between the time series of returns, and because our model captures nonlinear dependencies that can be parameterized by piecewise continuous functions.

We use the Lévy-area between each pair of stock returns to find nonlinear lead-lag relationships. We show that when the return of one stock is modeled as a continuous function of the previous return of another stock, the Lévy-area between the returns of these two stocks reflects the direction and strength of the lead-lag relationship between them. For more detail on Lévy-area in machine learning, data streams, and quantitative finance, see Lyons (2014); Gyurkó et al. (2013); Chevyrev and Kormilitzin (2016). In the machine learning and statistics literature, there are data-driven methods for time series lead-lag detection. One of the most common and intuitive methods is to compute lagged correlation or cross-correlation between pairs of time series. Cross-correlation is a class of lagged correlation coefficients between two time series. Numerous works in finance use this method. For example, Chan (1992) studies intra-day lead-lag relationships between the cash market and stock index futures market. In the computer science literature, Sakurai et al. (2005) use lagged correlation to develop the so-called data-stream mining method. For univariate regression models, the R^2 value of a regression is equivalent to the square of the correlation coefficient between the dependent variable and the independent variable after normalizing the variables. Therefore, using cross-correlation as lead-lag detection method is equivalent to conducting univariate linear regressions and using R^2 as a measurement criterion of

lead-lag relationships.

On the other hand, works including those of Badrinath et al. (1995); Chordia and Swaminathan (2000); Hou (2007); Frazzini and Cohen (2008); Menzly and Ozbas (2009) use the R^2 of the Granger causality test to verify the direction of the lead-lag relationships they study. While it is easier to establish statistical significance with the Granger causality test, directly using cross-correlation better aligns with the scope of this paper because we do not assume the direction of lead-lag relationships and we wish to explore the dynamics of lead-lag relationships. Another popular method for lead-lag detection is the spectral method in Hause (1971). If two time series are related by a lead-lag relationship, then the cross-spectral density function will exhibit a peak at a certain frequency corresponding to the lag between the two series. This method is most useful when the time series are driven by cyclical or periodic patterns. However, in finance, the spectral method is less useful and not frequently employed for lead-lag detection because stock returns exhibit more complex and often nonlinear relationships.

Finally, in our study of the slow information diffusion hypothesis, we are the first to study lead-lag relationships across various sampling frequencies. Most papers study lead-lag relationships at a monthly frequency (Badrinath et al. (1995), Menzly and Ozbas (2009), Frazzini and Cohen (2008), Parsons et al. (2020), Huang et al. (2022)), and others study weekly lead-lag relationships (Lo and MacKinlay (1990), Chordia and Swaminathan (2000), Hou (2007)). Few papers study daily lead-lag relationships, e.g., Badrinath et al. (1995), Brennan et al. (1993), and Chordia and Swaminathan (2000). Hou (2007) argues that daily data may induce problems caused by non-synchronous updates of price information, which is mainly caused by the low liquidity of some stocks. In our study, we only use relatively large market capitalization stocks; therefore, we minimize instances

of non-synchronous price updates. The scarcity of studies employing daily data may be attributed to the prevalence of previous research that employs firm characteristics, which are available only at weekly or monthly intervals. With our approach, one can investigate frequencies from daily to monthly because our method only requires asset prices, which are available at much higher frequencies than that of firm characteristics data.

The remainder of the paper is organized as follows. Section 4.2 presents the mathematical setup to identify lead-lag relationships and provides the main mathematical models we use. Section 4.3 describes the data and discusses results. Section 4.4 provides robustness checks and considers alternative specifications of numerical experiments. Section 4.5 concludes and the appendix collects proofs and results.

4.2 Problem Setting

Here, we identify and verify lead-lag relationships in a set of stocks in two steps: identify potential leaders and followers, and construct a lead-lag portfolio to verify the lead-lag relationship.

4.2.1 Model setup and pairwise lead-lag relationships

Consider an economy with assets A and B whose returns are

$$R_{A,t} = \beta_A \, f_t + \epsilon_{A,t} \,, \tag{4.1a}$$

$$R_{B,t} = \beta_B \, f_t + \tilde{\beta}_B \, f_{t-\ell} + \epsilon_{B,t} \,, \tag{4.1b}$$

where f_t is a latent risk factor that is independent and identically distributed (i.i.d.) across time, with mean zero, the error terms $\epsilon_{A,t}$ and $\epsilon_{B,t}$ are i.i.d., and the lag $\ell \geq 0$.

In (4.1), the exposure of both assets to a contemporaneous common source of information f is linear, and asset B is also exposed to previous values of the factor f due to slow information diffusion. For conciseness, we assume that the sign of all risk exposures is the same. Also, for ease of presentation, we assume that the variance of returns of assets A and B are both standardized to 1.

In practice, because the risk factor f is latent, one observes a lead-lag relationship between the two assets because the returns of asset B are often positive/negative following positive/negative returns of asset A.

Next, one can compute the cross-correlation of returns

$$\begin{aligned}
\rho_{AB}(k) &= \mathbb{E}(R_{A,t-k}\, R_{B,t}) \\
&= \frac{\mathrm{Cov}(R_{A,t-k}, R_{B,t})}{\sqrt{\mathrm{Var}(R_{A,t-k})\,\mathrm{Var}(R_{B,t})}} \\
&= \begin{cases} \beta_A\, \tilde{\beta}_B & \text{if } k = \ell\,, \\ \beta_A\, \beta_B & \text{if } k = 0\,, \\ 0 & \text{if } k \neq 0 \text{ and } k \neq \ell\,, \end{cases}
\end{aligned}$$

to determine which asset is the leader and to develop tests for the null hypothesis that $\ell = 0$ (i.e., there is no lagged exposure to the risk factor f, hence no lead-lag relationship) against the alternative hypothesis that $\ell > 0$.

Two of these tests in our study below are

$$C^1(R_{A,t}, R_{B,t}) = \underset{\ell \in [-T,T]}{\arg\max}\; Corr(R_{A,t},\, R_{B,t-\ell})\,, \tag{4.2}$$

$$C^2(R_{A,t}, R_{B,t}) = \pm \frac{1}{T} \max \left\{ \sum_{\ell \in [-T,0]} Corr(R_{A,t},\, R_{B,t-\ell}) \,, \sum_{\ell \in [0,T]} Corr(R_{A,t},\, R_{B,t-\ell}) \right\} . \tag{4.3}$$

Here, $Corr(R_{A,t},\, R_{B,t-\ell})$ is the Pearson correlation between the returns $R_{A,t}$ and $R_{B,t-\ell}$, and T is the lookback window. We refer to these two tests as C^1 and C^2, respectively.

Each of these two tests has their own shortcomings. For example, if $\beta_B > \tilde{\beta}_B$, the test statistic $C^1(R_{A,t}, R_{B,t}) = 0$, so the probability of reporting false negatives is high. Thus, the power of the C^1 test is low when the exposure to both the contemporaneous factor f_t and the lagged factor $f_{t-\ell}$ of the follower asset B is substantial. On the other hand, C^2 cannot identify cases when there is no lead-lag relationship by design; therefore, C^2 will report many false positives, which makes the size of this statistical test small.[5]

To address the shortcomings of methods based on cross correlations, we consider the particular case of one period lag $\ell = 1$. Our method employs the Lévy-area between pairwise stock returns to identify the leader and the follower (see Appendix A.2 for more details).

For a two-dimensional random process $X_t = (X_t^i, X_t^j)$, consider its coordinate iterated integrals over a time interval (s, t), defined as

$$S(X)_{s,t}^{i,j} = \int_{s<a<t} \int_{s<b<a} dX_b^i \, dX_a^j \,, \tag{4.4}$$

which we use to compute the Lévy-area

$$A_{i,j}^{Lévy} = \frac{1}{2} \left(S(X)_{s,t}^{i,j} - S(X)_{s,t}^{j,i} \right) . \tag{4.5}$$

[5]Here, the term size and power refer to the probability of a statistical test to report Type II and Type I error, respectively.

Proposition 1 *Assume returns are as in* (4.1). *Then, if* $\ell = 1$, *the Lévy-area* $A^{Lévy}_{A,B}$ *between the returns of assets* A *and* B *satisfies* $\mathbb{E}\left(A^{Lévy}_{A,B}\right) = \mathbb{E}\left(-A^{Lévy}_{B,A}\right) = \beta_A\,\tilde{\beta}_B$, *and if* $\ell \neq 1$, *then* $\mathbb{E}\left(A^{Lévy}_{A,B}\right) = 0$.

From Proposition 1, we know that when the lag in the exposure to the risk factor f is 1 unit of time, the Lévy-area can, on average, distinguish if a lead-lag relationship between the pair of assets exists. Practically, by adjusting the frequency at which one samples data (e.g., using daily or monthly returns), the lag $\ell = 1$ we use in our test can extend to other lengths of time without loss of generality. While only studying the case $\ell = 1$ does not consider cases where the lag size differs across stocks, our approach is consistent with the literature. In the literature, the approach of fixing the length of the lag is standard, e.g., Lo and MacKinlay (1990) consider only fixed lag lengths, and the length does not change across all pairs of leaders and followers.

A key advantage of the Lévy-area between returns is that it detects both linear and nonlinear lead-lag relationships. Consider the model

$$R_{A,t} = \beta_A\, f_t + \epsilon_{1,t}\,, \tag{4.6a}$$

$$R_{B,t} = \beta_B\, f_t + \tilde{\beta}_B\, g(f_{t-\ell}) + \epsilon_{2,t}\,, \tag{4.6b}$$

where g is a nonlinear function, so the exposure of the follower to the latent factor f is nonlinear. Then, if $\mathbb{E}(x\, g(x)) \geq 0$ for a random variable x with 0 mean and unit variance,[6] we prove the following theorem.

Theorem 1 *Assume returns are as in* (4.6). *Then, if the lag is* $\ell = 1$ *unit of time, and*

[6]This assumption covers a large class of functions g including all monotonic functions and hence all linear functions.

if $\mathbb{E}(f_t\, g(f_t)) \geq 0$*, the Lévy-area* $A^{Lévy}_{A,B}$ *between the returns of assets* A *and* B *satisfies* $\mathbb{E}\left(A^{Lévy}_{A,B}\right) = \mathbb{E}\left(-A^{Lévy}_{B,A}\right) = k_{g,f}\,\beta_A\,\tilde{\beta}_B$ *for some finite positive value* $k_{g,f}$ *dependent on the function* g *and the distribution of* f*, and if* $\ell \neq 1$*, then* $\mathbb{E}\left(A^{Lévy}_{A,B}\right) = 0$*.*

For a proof, see Appendix A.3.

The above theorem shows that one can, on average, correctly identify the leader in a pair of assets when the risk factor has a nonlinear effect on the dynamics of the returns of the follower asst. Also, the theorem shows that the strength of the lead-lag relationship is proportional to the Lévy-area between the pair of assets. Thus, the Lévy-area enables us to study a larger class of lead-lag relationships where the exposure of the follower asset to risk factors may be linear or nonlinear.

4.2.2 Lead-lag relationships among many stocks

Now, consider N stocks. We sort the stocks from "most likely to be a leader" to "most likely to be a follower" with a criterion that measures how much each stock leads other stocks over a lookback window. The leaders are stocks that likely react the fastest to risk factors, and the followers are likely the ones that react the slowest to common risk factor that also affect the leader.

We proceed as follows. For each stock, compute its pairwise lead-lag test statistics against all other stocks. Next, take the average of the pairwise lead-lag test statistics and sort stocks according to this value from higher to lower, where we regard stocks with large average pairwise lead-lag test statistics as leaders. At each time t, we assume that those that lead many stocks in the past will continue to lead many stocks in the near future. That is, we assume that lead-lag relationships exhibit short-term momentum.

To implement the above procedure, we use the pairwise lead-lag test statistics to populate an N by N matrix. In this matrix, which we denote $\mathbf{C}$, entry $C_{i,j}$ is the test statistic on the null hypothesis that there is no lead-lag relationship between stocks i and j. Recall that we regard stocks with large average pairwise lead-lag test statistics as leaders. This is equivalent to sorting stocks by the column mean of this matrix.

Many lead-lag measurements proposed in the literature are analogous to our approach. For example, sorting stocks by market capitalization in Lo and MacKinlay (1990) is equivalent to using the pairwise difference in their market capitalization as a test statistic for lead-lag relationships. In our study, instead of assuming some firm characteristic indicates leaders and followers, we directly measure lead-lag relationship in the last lookback window and use this measurement as our "best guess" of lead-lag relationships in the near future.

Therefore, the mean C_i of the column corresponding to stock i in matrix $\mathbf{C}$ is positively correlated to the propensity of a stock being a leader among N stocks. Here, C_i is a measurement of how much stock i leads other stocks over the last lookback window. From this perspective, C_i is analogous to a firm characteristic (e.g., market cap, turnover rate, etc.) that the literature assumes to be positively correlated to the likelihood of a stock being a leader. Our approach is more direct because we explicitly measure lead-lag relationships, and we offer a more flexible and dynamic way of identifying leaders and followers.

Sorting stocks by their mean value of pairwise test statistics is not the only way to identify leaders and followers. Appendix A.4 introduces three alternative methods to sort stocks based on the matrix of their pairwise test statistics; we use these alternative methods in our robustness analyses.

4.2.3 Interpreting lead-lag relationships

The lead-lag relationships we detect are easy to interpret. On each day t, our method measures the leadership strength of each stock over the remaining $N-1$ stocks during the past lookback window. With this measurement of overall leadership, we build a linear model to understand how much of this leadership is explained by the lead-lag relationships studied in the literature. Consider the regression model

$$C_i = \sum_i \beta_i \, \mathbf{f}_{B,i} + \epsilon_B \, .$$

Here, the left-hand side of the equation is the column average of the lead-lag matrix,[7] and the right-hand side of the equation is a linear model that uses the (standardized) firm characteristics to explain why a stock was a leader in the previous lookback window. The firm characteristics include market capitalization, rolling average of turnover ratio, analyst coverage, and information discreteness (Huang et al. (2022)).

We perform one regression per day, and we compute the average coefficients and R-square to understand how much of the short-term lead-lag relationship we detect can be explained by well-known lead-lag relationships such as size and liquidity.

4.2.4 Constructing lead-lag portfolios

After sorting the assets from most likely to be a leader to most likely to be a follower, we use the raw returns of the leaders to decide whether to buy or to sell the followers, and we construct a portfolio to evaluate the economic significance of the lead-lag relationship. We use the portfolio to understand if short-term (daily) lead-lag relationships exists. Ad-

[7] Recall, this measures how much one stock, on average, leads other stocks.

ditionally, because we focus on short-term lead-lag relationships, we simultaneously test the persistence of lead-lag relationships, i.e., if stocks that lead others in the previous 30 to 60 days tend to continue be leaders in the next day.

At each time t, let the top m stocks represent the set of leaders n_l^t and let the bottom n stocks represent the set of followers n_f^t in the sorting. Then, the raw returns of the m stocks at time $t-\ell$ should provide an indication of the raw returns of the n stocks at time t. Furthermore, because we expect the predictability of the m leaders on the n followers to be stronger than that of the m leaders on the US equity market, the sign of the raw returns of the m leaders at time $t-\ell$ should be, statistically, the same as the sign of the difference between the raw return of the n follower stocks and the market return R_{mkt}.

Therefore, we build a portfolio to evaluate if the lead-lag relationship we detect is economically significant; see Algorithm 1 and Figure 4.1a. The algorithm uses the previous raw returns of the leaders as a signal to construct an equal-weight portfolio with the followers. Here, R_{t-1}^{Leader} and $R_t^{Follower}$ are the average raw returns of an equal-weighted portfolio with the leader stocks at time $t-1$ and the portfolio of the follower stocks at time t, respectively. When the average of the raw returns of the leaders is positive, the algorithm buys the followers, and when the average of the raw returns of the portfolio of leaders is negative, the algorithm sells the followers. In step 3 of the algorithm, one trades the market to finance the portfolio. For example, when the average of the raw returns of the leaders is negative at time t, one short sells the followers and buys the market at time $t+1$. Therefore, one way of interpreting the performance of the portfolio is by how much more the leaders predict the followers than they predict the US equity market.

An alternative to this approach is to trade the spread between the top and bottom quan-

Algorithm 1 Constructing a lead-lag portfolio

Input: Raw asset returns $R_{1,1:t}, \cdots, R_{N,1:t}$, Raw market return $R_{mkt,1:t}$
Parameter: Look-back window w, number of leaders m, number of followers n; $m+n \leq N$
while $t \geq w$ **do**
  1. Compute skew-symmetric matrix $C_{t-w:t}$, whose entries are pairwise lead-lag scores **2.** Sort by column average of $C_{t-w:t}$, pick top m assets as leaders, bottom n assets as followers **3.** Compute R_{t-1}^{Leader} and $R_t^{Follower}$
  if $R_{t-1}^{Leader} \geq 0$ **then**
    $R_t^{Portfolio} = R_t^{Follower} - R_{mkt,t}$
  else
    $R_t^{Portfolio} = R_{mkt,t} - R_t^{Follower}$

tiles of stocks after sorting, i.e., to finance the portfolio with the leaders instead of the market. For example, if at time t the average of the raw return of the portfolio of leaders is negative, one short sells the followers and buys the leaders at time $t+1$. This alternative approach assumes that the predictability of the leader on the follower is greater than the autocorrelation of the leader. Compared with this alternative, our approach offers two primary advantages. First, the performance of our portfolio is interpretable; it measures the ability of the portfolio of leaders to predict the portfolio of followers instead of predicting the market. Second, it is more feasible to trade exchange-traded funds such as the SPY to finance the lead-lag portfolio. In particular, when short selling, it is more feasible to short sell the SPY ETF than an entire portfolio of leaders. In the context of financial time series; our approach to consider the performance of the portfolio is more suitable than those in the literature that apply linear regression models and observe R^2 values because the variance of forecasts can heavily influence the predictive R^2, see Kelly et al. (2022).

Some studies use characteristics including firm sector or the location of a firm's headquarter to group stocks into blocks. Next, they sort the stocks into leaders and followers within each of these blocks and construct lead-lag portfolios within each block of stocks as above, see, e.g., Hou (2007). To compare our approach with those in the papers that classify stocks

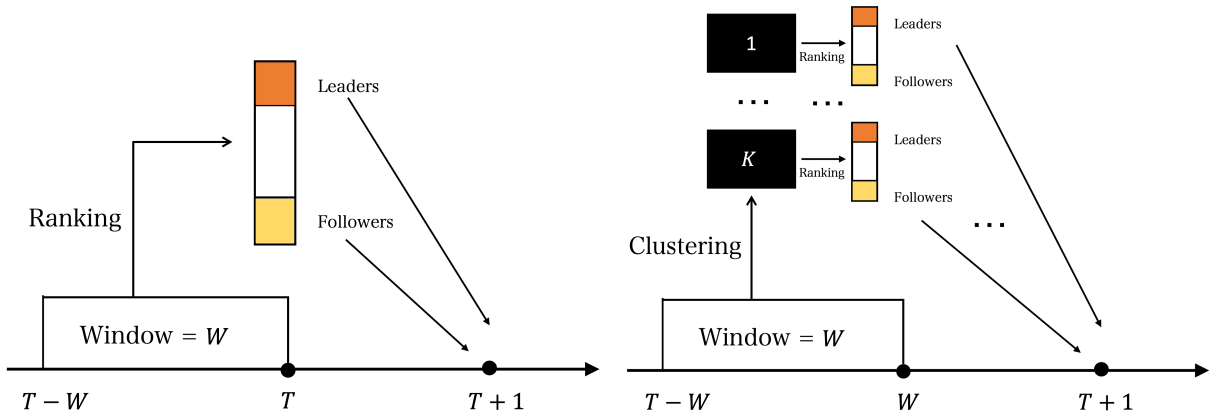
into groups, we extend our methodology to include data-driven algorithms that partition stocks into groups. Specifically, we use state-of-the-art data-driven clustering algorithms, including spectral clustering (Ng et al. (2001)) and Hermitian clustering (Cucuringu et al. (2019b)) that use information in the lead-lag matrix $\mathbf{C}$ to group stocks. Financially, these algorithms rearrange assets into groups via pairwise characteristics. We apply these clustering algorithms to the skew-symmetric lead-lag matrix $\mathbf{C}$ obtained with Algorithm 1, and employ the resulting clusters to construct our portfolios. This extension to our approach allows meaningful comparisons with the lead-lag relationships identified in the literature, and the approach enables comparisons between lead-lag relationships we detect and the relationship induced by pre-defined clusters such as sectors. See Figure 4.1b for a visual illustration of our clustering method.

From now on, we refer to the portfolio illustrated in Figure 4.1a as the “global" lead-lag portfolio, and we refer to portfolios illustrated in Figure 4.1b as the “clustered" lead-lag portfolio. In the global lead-lag portfolio, we rank the assets into most likely leaders and most likely followers, and we use the returns of the leaders to predict the returns of the followers. In the clustered portfolio, we first cluster the stocks into K blocks and construct lead-lag portfolios within each block as in Algorithm 1. Initially, we assume zero execution costs, and in Section 4.4.1 we include trading fees and frictions.

4.3 Results

4.3.1 Data

Asset prices and firm characteristics data are from the Center of Research in Security Prices (CRSP) daily returns database. The sample period is from July 1963 to December



(a) Global portfolio (b) Clustered portfolio

Figure 4.1: Global portfolio: Assets are ranked into most likely leaders to most likely followers, and the returns of the leaders are used to predict the returns of the followers, see Algorithm 1. Clustered portfolio: Assets are clustered into K blocks, and K lead-lag portfolios are constructed within each block as in Algorithm 1.

2022. To be consistent with standard literature practices, we include NYSE, Amex, and NASDAQ stocks. For each trading day, to ensure that the trading positions we take are feasible, we use the stocks in the top 25 percentile of market capitalization, which is given by the product of the price of the stock at the end of the day and the number of shares outstanding.

For all frequencies (e.g., bidiurnal, weekly, bi-weekly, monthly, etc.), returns are computed from daily returns as the cumulative product of one plus the end-of-day return of the assets.[8] For example, for a calendar week with five trading days and raw returns $R_1, R_2, \cdots, R_5$, the raw return of the week is $R_{week} = \prod_{i=1}^{5}(1+R_i)-1$, and the raw return of the week is standardized to compute the Lévy-area and cross-correlation between pairs of stock returns.

To study the composition of portfolios, we also include industry classification data. The industry classification information links each firm to a single, non-overlapping Fama–French 12 industries by its SIC code. The industries are (1) nondurables, (2) durables, (3)

[8]Daily return for stock n at day t is computed by purchasing the stock on the most recent time before t when the stock had a valid price, see CRSP.

manufacturing, (4) energy, (5) chemicals, (6) business equipment, (7) telecommunications, (8) utilities, (9) shops, (10) healthcare, (11) finance, and (12) other.

4.3.1.1 Summary Statistics

Tables A.2 to A.6 in Appendix A.5 provide summary statistics for the set of assets used in this paper, broken down by decade and by Fama–French sector labels for stocks. Table A.2 shows the average number of stocks that are included in our study from 1963 to 2022, split by decade and industry. Recall that on each trading day we construct the portfolio with the stocks in the top 25 percentile by market capitalization. In addition, to compute lagged correlation and Lévy-area, we use the stocks that do not have missing returns in the previous 60 trading days. From left to right in the table, we see an increasing trend in the number of stocks included in our study, with an average of 143 in the 1960s to 593 in the 2010s. However, the increase in number of stocks is not monotonic and stocks are distributed unequally across industries. For nondurables, durables, manufacturing, and utilities, we observe a monotonic decrease in the number of stocks between the 1990s and the 2010s. This is likely correlated to the evolution of the global economy which caused the number of companies in these four industries to decline in the US. On the other hand, for business equipments, telecommunications, shops, healthcare, and finance stocks, we observe a peak in the number of stocks in the 2000s. We also observe the same trend in the total number of stocks. The count of stocks in several industries peaks in the 2000s and declines in the 2010s; this observation coincides with the 2008 financial crisis.

Table A.3 shows the mean and standard deviation of stock prices, and Table A.4 shows the mean and standard deviation of stock returns for the stocks we use to build portfolios. Overall, we observe a strong upward trend in the mean and the volatility of stock prices

over time, but the average daily return and volatility of stock returns do not undergo considerable changes. The mean and volatility of stock prices in the finance sector are much higher than those in other sectors. This is primarily due to the high unit price of Berkshire Hathaway (BRK.A), which lists at approximately 500,000 dollars per share at the time this paper is written.[9] Meanwhile, for stock returns, we observe an increase in volatility and decrease in mean returns during the 2000s. This is likely correlated to the dotcom bubble and the 2008 financial crisis. We observe this trend across all Fama–French industries.

Tables A.5 and A.6 show the mean and the standard deviation of the volume and of the turnover ratio for the stocks, respectively. There is a clear upward trend in both volume traded and turnover ratio (defined as the quotient of the daily volume of shares traded to the shares outstanding) for stocks, indicating an increase in trading activity. In particular, average daily volume for each stock increased by over five times from the 1970s to the 1980s, over three times from the 1980s to the 1990s, and over five times from the 1990s to the 2000s. Similarly, the turnover ratio of stocks nearly doubles between 1970s and 1980s, 1980s and 1990s, and 1990s and 2000s. This is likely correlated to the emergence of algorithmic trading in the 1980s and the 1990s, and to the development of high-frequency trading in the 2000s.[10] Another possible reason for the change in trading volume and turnover ratio is the decimalisation of the US market in 2001, which reduced spreads and thus lowered execution costs, see Bessembinder (2003). On the other hand, the average turnover of stocks does not undergo large changes from the 2000s to the 2010s

[9]The mean price of stocks in the finance sector is similar to that of other sectors after removing BRK.A.

[10]Program trading became widely used in the 80s in trading between the S&P 500 equity and futures markets (McGowan (2010)). Electronic trading venues known as electronic communications networks (ECNs) emerged in the late 1990s and allowed trading of financial products outside of the traditional stock exchanges (McGowan (2010)). In the US, market share of high-frequency trading increased from less than 20% in 2005 to around 60% in 2009 (Zaharudin et al. (2022)).

(13.6 and 13.2 respectively). Hence, the increase in share volume traded from the 2000s to the 2010s is driven by the increase in shares outstanding. The observed trends in market liquidity are consistent across all industries.

4.3.2 Evidence of daily lead-lag relationship

In this section, we build lead-lag portfolios to investigate lead-lag relationships. Our goal is to test if stocks that lead other stocks in the previous days tend to continue to be leaders in the following trading day. We rank the stocks and rebalance the portfolios at a daily frequency because this is the most granular frequency at which one can compare our results with the benchmarks. The benchmark portfolios use firm characteristics that are available at daily frequencies, e.g., market capitalization and trading turnover. We employ Algorithm 1 to construct the portfolios and provide empirical evidence that the data-driven lead-lag portfolios outperform the baseline lead-lag portfolios.

For our data-driven lead-lag portfolio, the Lévy-area and cross-correlation between pairwise stock returns are computed over a 60-day lookback window; the stocks we include on each trading day are the top 20% (i.e., leaders) and bottom 20% (i.e., followers); and the market return is that of the SPY ETF. Recall that we consider the top quantile of all stocks in terms of market capitalization to construct our portfolio. Appendix A.6 presents the results for alternative choices of look-back windows and various choices of the number of leaders and followers.[11] Additionally, when comparing our portfolios with those in the literature that group stocks into small blocks, we choose spectral clustering to form 12 clusters, which is the same number of Fama–French industries. Appendix A.6 presents the

[11]Most of the literature splits stocks into quantiles. They use the top quantile as leaders and bottom quantile as followers. In Appendix A.6, we use the top 40% of the stocks as leaders and the bottom 40% of stocks as followers.

results for alternative choices of clustering methods.

To compare our findings with those in Lo and MacKinlay (1990), Chordia and Swaminathan (2000), and Hou (2007), which are our main benchmarks, we adapt their lead-lag portfolio construction procedures to align with that in Algorithm 1. Specifically, for Lo and MacKinlay (1990), we sort stocks by market capitalization (i.e., the product of the shares outstanding on the previous trading day and the previous day's price), and we use the top 20% stocks as leaders and the bottom 20% stocks as followers to construct the portfolio. We use the SPY ETF to finance all portfolios, as in our data-driven lead-lag portfolios. For Chordia and Swaminathan (2000), we use the same steps as above, except that instead of market capitalization, we use turnover to rank the stocks from leaders to followers. Similarly, for Hou (2007), we group stocks into their respective Fama–French industries and sort them by market capitalization within each group. Next, we use the approach of Lo and MacKinlay (1990) to construct and finance lead-lag portfolios for each industry; therefore, on each trading day, the lead-lag portfolio for Hou (2007) consists of 12 portfolios with zero net initial capital (i.e., hold long and short positions of the same dollar value on assets in each portfolio initially). Appendix A.4 provides the results for alternative choices of ranking algorithms, and Appendix A.6 provides results for alternative choices of the percentage of stocks that are considered leaders and followers.

Table 4.1 shows the main results for various lead-lag portfolios. Panel A reports the performance of the data-driven lead-lag portfolios compared with those of Lo and MacKinlay (1990) and Chordia and Swaminathan (2000), denoted by "Market Cap" and "Turnover", respectively. Panel B shows the difference in performance between clustered lead-lag portfolios from the data-driven methods and the portfolio from Hou (2007), which is denoted "Industry". In both panels, we evaluate the portfolios with five criteria: annualized return,

Table 4.1: Performance of various lead-lag portfolios

Panel A: Global Lead-Lag Portfolios					
	Ann. Return (%)	Return (bps/day)	Daily Vol. (%)	Sharpe	Max Drawdown (%)
Max Cross-Cor	19.69	7.14	0.47	2.37	16.90
Avg Cross-Cor	27.97	9.79	0.70	2.21	28.64
Lévy-area	24.87	8.82	0.59	2.38	24.67
Market Cap	6.15	2.37	0.50	0.75	63.40
Turnover	7.60	2.91	0.34	1.37	13.62
Panel B: Clustered Lead-lag Portfolios					
	Ann. Return (%)	Return (bps/day)	Daily Vol. (%)	Sharpe	Max Drawdown (%)
Max Cross-Cor	14.52	5.39	0.38	2.23	18.21
Avg Cross-Cor	19.01	6.91	0.49	2.23	23.49
Lévy-area	17.50	6.40	0.42	2.43	12.28
Industry	15.75	5.81	0.43	2.18	42.29

daily return, daily volatility, portfolio Sharpe ratio, and maximum drawdown of the portfolio. The return of the portfolio reflects the predictive power of the lead-lag relationship; volatility and Sharpe ratio show the risks of lead-lag portfolios; and maximum drawdown indicates when lead-lag relationships temporarily break down.

The data-driven lead-lag portfolios outperform the benchmarks on almost all categories except for volatility, see Panel A in Table 4.1. The daily returns for the portfolios range from 2.37 basis points (bps) per day for the Market Cap portfolio to over 9 bps per day for the average cross correlation portfolio. The Sharpe ratios of our data-driven portfolios are all above two and the annualized returns are above 20%. In particular, the annualized return of the Lévy-area portfolio is 24.87% over the span of 60 years. Additionally, we compute criteria that are sensitive to risk to understand further the performance of the portfolios. The Sharpe ratios of the data-driven portfolios are higher than those of the two benchmark portfolios based on firm-characteristic-driven lead-lag identifications. This shows that the methods that rank stocks based on pairwise lead-lag relationships capture economically significant lead-lag relationships. In particular, the Sharpe ratio of the Lévy-area portfolio is the highest, and the annualized return of the Lévy-area portfolio is the

second highest among our three data-driven methods reported in Panel A of Table 4.1. In Appendix A.6, we show that the Sharpe ratio and annualized return of the Lévy-area portfolio are also the highest when we propose an alternative choice of proportion of leaders and followers to construct the portfolios. Overall, our results show that the Lévy-area portfolio outperforms the cross-correlation-based methods.

Panel B of Table 4.1 compares the performances of various clustered lead-lag portfolios. The performances of the Max Cross-Cor, Avg Cross-Cor, Lévy-area, and Industry portfolios are economically significant, with returns ranging from 5.39 to 6.91 bps/day. The data-driven portfolios Max Cross-Cor, Avg Cross-Cor, and Lévy-area, outperform the Industry portfolio for most criteria, including annualized return and Sharpe ratio. In particular, compounded return of the Lévy-area portfolio is 17.5%, with a relatively low volatility of 0.42%. In terms of criteria that account for risk exposures, the data-driven portfolios also perform better than the baseline Industry portfolio, with higher Sharpe ratios and lower historical maximum drawdowns. In particular, the Sharpe ratio for the Lévy-area portfolio is 2.43, which is at least 0.2 higher than that of any other portfolio. This suggests that the data-driven methods used to construct the lead-lag portfolios capture economically significant lead-lag relationships, with the Lévy-area method being particularly effective.

4.3.3 Understanding lead-lag relationships

On each day, we run a cross-sectional regression to understand how much of the short-term lead-lag effects we detect in the market can be explained by effects previously known in the literature. We assess if the coefficients we obtain in the regression are statistically significant by computing their t-statistics across each regression under the null hypothesis that the coefficients are zero.

Table 4.2: Daily regression on rank of stock

Coefficient and R-square of regression the cross-sectional percentage rank of each stock according to C^1, C^2, and Lévy-area over the past 30 days against the cross-sectional percentage rank of stocks' market capitalization, number of analysts covering the stock, 30-day median turnover ratio, and Information Discreteness indicator. All results are reported in percentages, and zero-mean t-statistics are reported in braces.

	Market Cap	Analyst coverage	30 day Turnover	Info. Discreteness	Avg. R2
Max Cross-Cor	6.15	−1.17	7.90	−1.99	2.93
	(7.54)	(−3.19)	(5.34)	(−4.38)	
Avg Cross-Cor	7.05	−1.34	7.58	−2.63	2.93
	(7.48)	(−3.79)	(5.97)	(−5.02)	
Lévy-area	3.94	4.45	−1.91	0.96	5.29
	(4.37)	(6.08)	(−0.64)	(2.02)	

From Table 4.2, short-term lead-lag relationships detected by C^1 and C^2 can be partially explained by large market capitalization, high liquidity, low information discreteness, and low analyst coverage. While the first three effects are consistent with discoveries in the literature, the last one is the opposite of what Brennan et al. (1993) discovered. On the other hand, the short-term lead-lag relationships discovered by Lévy-area can be partly explained by high market capitalization, high analyst coverage, and high information discreteness, whereas liquidity does not seem to explain lead-lag relationships. The effect of size and analyst coverage are consistent with those discovered in the literature, but the effect of information discreteness is in the opposite direction compared with those found in the literature.

Correlation between the cross-sectional ranks of these firm characteristics is unlikely to explain the discrepancy we observe between what the lead-lag relationship explains and the results in the extant literature. No pair of firm characteristics we include in this regression analysis have correlations larger than 5% in magnitude.

One possible explanation for the discrepancies is that the lead-lag relationships we observe are on a daily level, whereas the literature mostly focuses on lead-lag relationships over

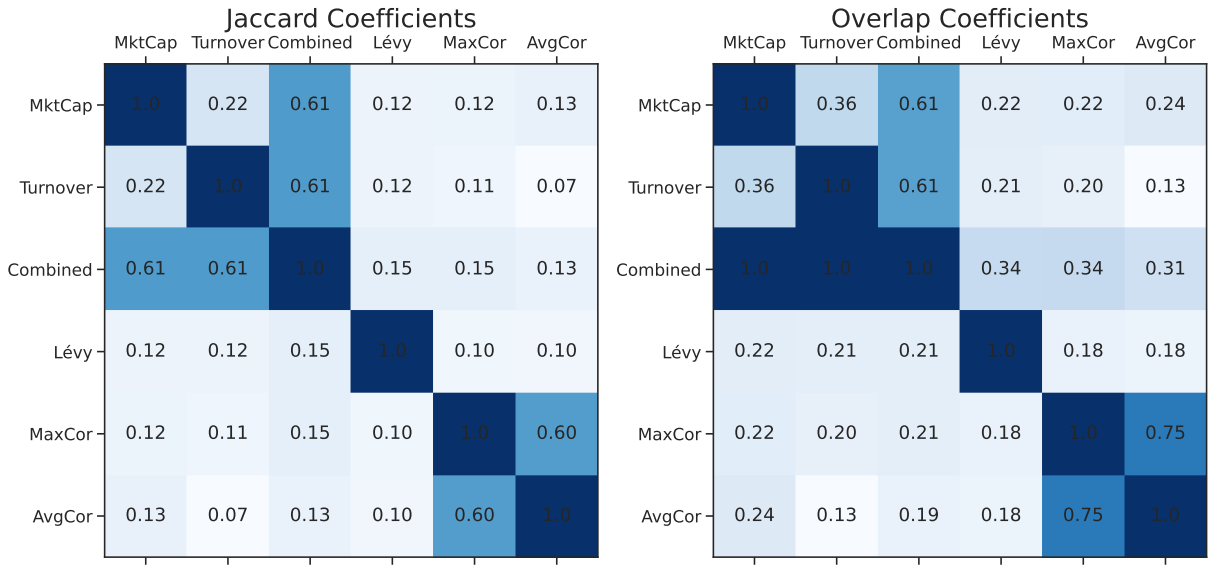
much longer horizons, e.g., weekly or monthly. Another possible reason why we observe different effects between C^1, C^2, and Lévy-area is that the Lévy-area detects potentially nonlinear lead-lag relationships, so it could be detecting latent effects that neither C^1 nor C^2 detects.

Overall, in the short-term, the lead-lag relationship between large and small size stocks seems to explain the market lead-lag effects consistently, whereas other effects the literature reports tend to have inconsistent impacts over the short-term.

4.3.4 Composition of lead-lag relationship

In this subsection, we study the composition of the data-driven lead-lag portfolios. We employ a three-fold approach to analyze the characteristics of leaders and followers and to compare the composition of the lead-lag portfolios with that of the benchmarks. First, we use the Jaccard score and the overlap coefficient to compute the similarity between data-driven portfolios and the benchmark portfolios. Second, we use the Adjusted Rand Index (ARI) to compute the similarity between the clusters formed for the data-driven portfolios and the underlying Fama–French sector labels of the stocks, see Hubert (1985). Third, we use a permutation test to assess if, for the leaders and followers, there is a statistically significant difference between the average market capitalization, volume, turnover, unit price, number of shares outstanding, and return. The goal of these three analyses is to understand and quantify the overlap between the constituents of the data-driven lead-lag portfolios found in our study and those reported in the literature.

In the first analysis, for each day and for each strategy, we compute the pairwise Jaccard score and the overlap coefficient for the stocks traded by each strategy (i.e., the followers identified by each trading strategy). Figure 4.2 shows the average Jaccard score and



(a) (b)

Figure 4.2: (a) Jaccard coefficients among lead-lag portfolios; a higher value of the Jaccard coefficient represents higher similarity (b) Overlap coefficients among lead-lag portfolios; a higher value of the overlap coefficient represents higher similarity.

Overlap coefficient. In the comparison, we include portfolios constructed based on Lo and MacKinlay (1990) and Chordia and Swaminathan (2000); recall they are denoted MktCap and Turnover, respectively, and our three data-driven methods are denoted Lévy, MaxCor, and AvgCor. To compare the similarity between our data-driven strategies and the benchmark, we also include a strategy which trades the union of stocks traded in MktCap and Turnover, which is denoted “Combined".

The Jaccard coefficient between the Combined strategy and the data-driven strategies is approximately 15%, and the respective overlap coefficients are around 35%, which shows that approximately 35% of the stocks traded in any one of the data-driven strategies is also traded by at least one of the baseline strategies that uses firm characteristics to identify leaders and followers, see Figure 4.2. A Jaccard score of 60% and an overlap coefficient of 75% between the MaxCor and the AvgCor portfolios show that these two data-driven benchmarks are similar. However, the low similarity between the correlation-

based portfolios and the Lévy-area portfolio shows that Lévy-area, which detects both linear and nonlinear lead-lag relationships, constructs a considerably different portfolio when compared with all other portfolios, both the data-driven and those proposed by the literature.

Table 4.3: Permutation test and adjusted Rand index analysis on Lead-lag portfolios

Panel A: Average ARI between lead-lag portfolios and the Fama–French industry labels

	Average ARI Spectral (%)	Average ARI Hermitian (%)
Max Cross-Cor	1.00	5.40
Avg Cross-Cor	3.17	3.19
Lévy-area	7.71	20.34

Panel B: Characteristics for leaders and followers in the Lévy-area portfolio

	Leader Avg Percentile (%)	Follower Avg Percentile (%)	Permutation test p-value
Market Cap	54.73	56.15	0.031
Volume	50.53	53.25	< 0.001
Turnover	38.82	41.28	0.025
Return	50.07	50.44	< 0.001
Price	56.65	55.31	0.103
Shares Outstanding	63.28	66.35	< 0.001

Next, we use spectral clustering and Hermitian clustering to compute the ARI between clusters formed with the data-driven portfolios and the underlying Fama–French industry labels. By doing this, we compare the clusters that we find with the Fama–French industry labels, used by Hou (2007), and therefore we study the overlap between our portfolios and those of Hou (2007). We split time from 1963 to 2003 into sub-periods and group the stocks into clusters to compare the clusters with the underlying stock sectors. First, we split time into 1000-day windows and form a new matrix with the sum of the lead-lag matrices in each window. In this step, we include stocks that are both in the top quantile in terms of market capitalization and are actively traded over the previous 1000 days. Next, we use spectral clustering and Hermitian clustering to split the new matrix into 12 clusters and use the ARI to compute the similarity between the formed clusters and the underlying sector labels. Panel A of Table 4.3 presents the results.

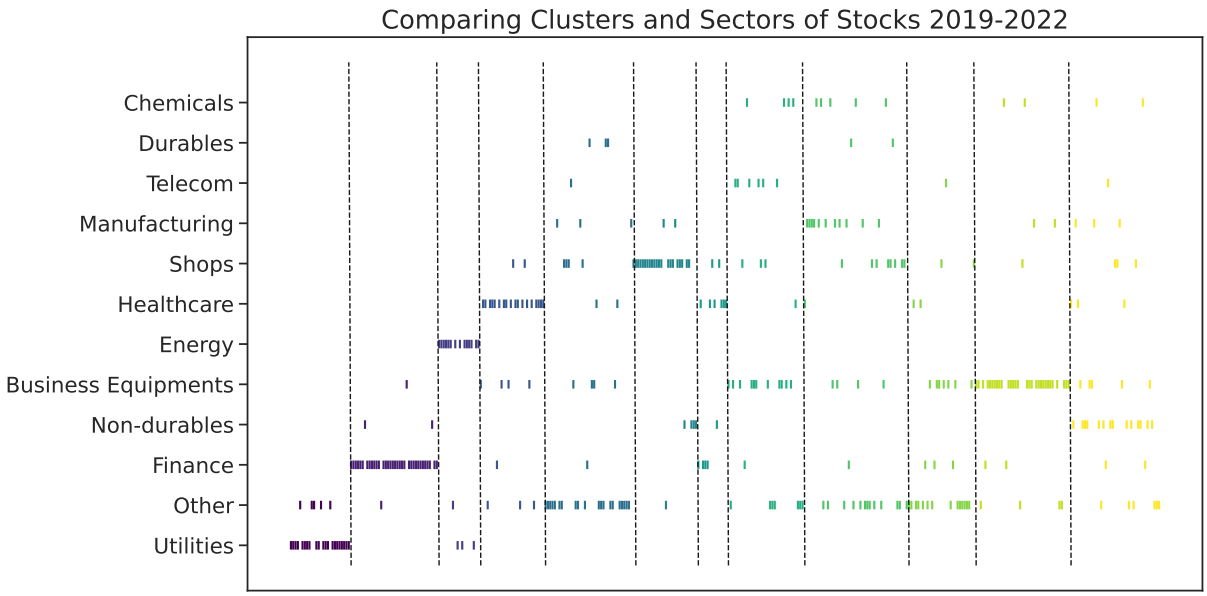


Figure 4.3: Clusters formed by Hermitian clustering on the Lévy-area matrices from 1 January 2019 to 31 December 2022. The area between black vertical dashes represents each cluster formed with Hermitian clustering. During this period, 376 stocks are traded.

Panel A in Table 4.3 shows that the similarity between the clusters for lead-lag portfolios and the Fama–French industry labels is low, with the exception of the Hermitian clusters for the Lévy-area portfolio. The data-driven portfolios arrange the stocks into groups that are almost unrelated to the industry labels; however, these data-driven clusters show stronger lead-lag relationships within each cluster than those within the Fama–French industries.

The Lévy-area portfolio with Hermitian clustering is the only case in panel A, Table 4.3, that is similar to the Fama–French 12 sectors with an ARI of approximately 20%. Figure 4.3 provides a visualization of clusters in the Lévy-area portfolio for the 1000-day window from 2019 to 2022. In the figure, each colored short dash represents a stock where the horizontal position represents cluster membership, and the vertical position represents sector membership. The black vertical lines represent the boundaries between each cluster.

Figure 4.3 shows that Hermitian clustering detects six clusters that are quite similar to

the underlying sector labels. The six clusters are mostly Energy stocks, Finance stocks, Business Equipments, Shops, Utilities, and Healthcare stocks, respectively. However, it is difficult to determine if the remaining six clusters and the sectors are similar.

The third analysis contains the permutation tests on the Lévy-area portfolio; see results in Panel B, Table 4.3. On each trading day, we compute the percentile score for each stock's average Lévy-area, market capitalization, turnover, volume, price, number of shares outstanding, and return for all stocks traded on that day. Next, we take the average of these percentile scores for each stock over its respective lifespan of active trading and sort all stocks that are ever traded (approximately 5000) by their average percentile of their Lévy-area. We take the top quantile as leaders and the bottom quantile as followers, and we use the permutation test to verify if there is a statistically significant difference between the mean of market capitalization, turnover, volume, and return of the leaders and the followers. Panel B in Table 4.3 shows that there is a statistically significant difference in average market capitalization, share volume traded, turnover, and return between the historical leader stocks and the follower stocks.

On average, stocks with a historical propensity to lead tend to have larger market capitalization, larger share volume traded, larger turnover, and high daily returns compared with stocks that are historically more often identified as followers. This difference is significant at the 5% confidence level for market capitalization and turnover, and it is significant at the 0.1% confidence level for volume and return. The differences in market capitalization, volume, and turnover of the leaders and the followers coincide with results in the literature; however, the extant literature does not discuss the differences in the returns for the leaders and followers. The discussion of why the leaders have higher historical average returns is beyond the scope of this paper and we defer this to future work. Results for

alternative choices of portfolios are included in Appendix A.7.

4.3.5 Sector lead-lag characteristics

In this section, we study if some sectors tend to contain more leaders and others contain more followers. Specifically, on each day, we use the ranking we infer from the lead-lag matrix to compute the percentile rank of each stock on each trading day. Then, we compute the average percentile rank for each Fama–French sector over time to understand the composition of each sector. We also inspect how the sector lead-lag identities evolve over time to understand if there is a change in the dynamics of sector-wise lead-lag relationships from 1963 to 2022.

Table 4.4: Average Percentile Rank of Stocks by Sector for Various Lead-Lag Matrix Constructions

Sector	**Lévy-Area (%)**	**MaxCor (%)**	**AvgCor (%)**
Manufacturing	47.6	50.1	50.1
Chemicals	48.5	50.4	50.4
Energy	49.4	50.4	50.4
Telecom	49.5	50.4	50.2
Healthcare	49.6	49.9	50.0
Utilities	49.8	50.7	50.5
Non-Durables	49.9	50.2	50.1
Shops	50.1	50.6	50.5
Finance	50.3	50.0	50.0
Other	50.4	49.8	49.8
Durables	51.1	50.1	50.2
Business Equipment	55.4	50.1	50.0

Table 4.4 presents the average percentile ranks of stocks in various sectors for different lead-lag matrix constructions from 1963 to 2022. The average percentile rank provides insights into the leader or follower status of each sector, with a lower rank indicating a higher proportion of leaders and a higher rank indicating a larger proportion of followers.

The results highlight that the inter-industry lead-lag relationships are predominantly sig-

nificant for the lead-lag relationship detected by the Lévy-area method. Sectors such as Manufacturing, Chemicals, and Energy exhibit much lower average percentile ranks, suggesting a higher proportion of leader stocks in these sectors. This coincides with the observation in Frazzini and Cohen (2008) that lead-lag relationships exist between customer-supplier linked firms with the supplier firms being leaders. On the other hand, the Business Equipments sector stands out with a notably higher average percentile rank, indicating a larger number of follower stocks within the sector. This implies that the stocks in Business Equipments tend to lag behind the other sectors. Overall, the Lévy-area method appears to detect inter-industry lead-lag relationships and finds sectors that tend to persistently contain leaders or followers.

In contrast, the cross-correlation-based lead-lag matrix constructions, including MaxCor and AvgCor, do not show significant differences in percentile rank across sectors. This suggests that the follower and the leader dynamics found by these methods are less persistent across sectors. The sectors demonstrate relatively similar average percentile ranks, indicating a balanced mix of leader and follower stocks within these sectors. This result likely shows that the cross-correlation-based methods do not account for the inter-industry lead-lag effects found by the Lévy-area method.

To further study the inter-industry lead-lag relationships detected by the Lévy-area method, we perform another experiment where we first compute the average lead-lag percentile rank of stocks in each sector for every year from 1963 to 2022 and then sort the sectors by their average lead-lag percentile rank from smallest to largest and record the rank of each sector, see Figure 4.4. We observe a significant shift in lead-lag identity across sectors over time. Business Equipments consistently lags other sectors from 1963 to around 2000, but does not rank at the bottom from 2000 onward; Finance exhibits strong leadership from

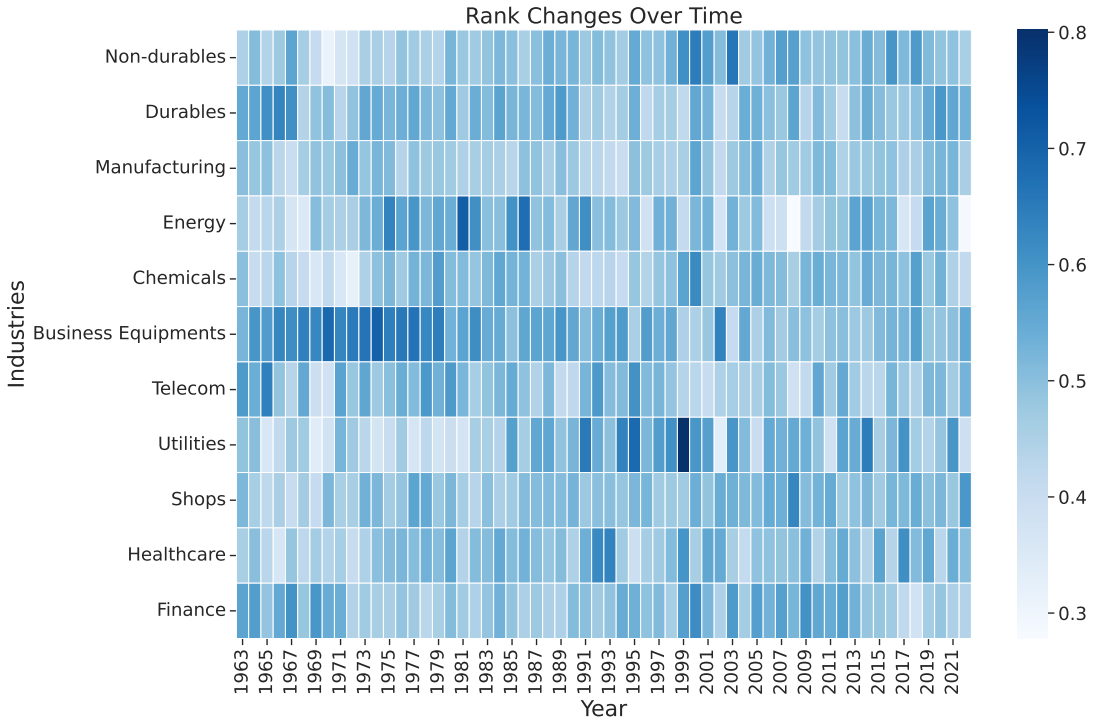


Figure 4.4: Change in sector rank of average lead-lag score for the Lévy-area lead-lag matrix construction from 1963 to 2022.

2017 onwards, but oscillates between leading and lagging prior to 2017. Utilities gradually shifts from a consistent leader in the 1970s to a consistent follower in the 1990s. Thus, the lead-lag relationships across sectors change, and the relationship constantly evolves over time.

4.3.6 Lead-lag relationship for various time scales

In this section, we investigate the lead-lag relationships for various time frequencies, which is a new contribution to the existing literature. Specifically, we explore the daily, bidiurnal, weekly, bi-weekly, tri-weekly, and monthly time frequencies to gain insight into the persistence of the lead-lag relationships. By doing so, we provide a more comprehensive understanding of these relationships and how their temporal dynamics evolve from 1963 to 2022.

4.3.6.1 Performance of lead-lag portfolios for various time scales

Most of the extant literature including Chordia and Swaminathan (2000), Hou (2007), Frazzini and Cohen (2008), Parsons et al. (2020), and Huang et al. (2022) hypothesize that lead-lag effects are consequences of the slow information diffusion. If this hypothesis holds, then lead-lag relationships should disappear as we reduce the time frequency with which we detect the relationship. Therefore, one would expect the performance of the lead-lag portfolios to decrease as the frequency with which one ranks leaders and followers decreases.

To analyze the lead-lag relationships at different frequencies, we aggregate returns from daily to other frequencies and compute the pairwise lead-lag scores as before. Recall that, to compute weekly lead-lag relationships, we compute weekly raw returns as $R_{t-5:t} = \Pi_{j\in[0,\cdots,5]}(1+R_{t-j}) - 1$ and relabel the return $R_{t-5:t}$ as R_{t_w}, where t_w is the w^{th} natural week of the year. Next, we normalize the computed weekly returns to feed into the steps in Algorithm 1 to detect lead-lag relationships.

Table 4.5 presents performances of the Lévy-area, MaxCor, and AvgCor portfolios for daily, bidiurnally, weekly, bi-weekly, tri-weekly, and monthly frequencies. Panels A, B, C in Table 4.5 show the results for Lévy-area, MaxCor, and AvgCor portfolios, respectively.

We observe an overall downward trend on portfolio performances as the rebalance frequency decreases from daily to monthly. The drop in annualized return and Sharpe ratio from the daily level to the bidiurnal level is the largest, and the subsequent performance drops from bidiurnal to lower frequencies is smoother and less pronounced. We also observe an overall upward trend in portfolio volatility as rebalance frequency decreases.

The results in Table 4.5 support the slow information diffusion hypothesis. The strength

Table 4.5: Performance of lead-lag portfolios for various frequencies

Panel A: Lévy-area portfolio for various frequencies					
	Ann. Return (%)	Return (bps/period)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Daily	24.87	8.82	0.59	2.37	24.67
Bi-Daily	11.02	8.30	0.76	1.23	24.17
Weekly	12.33	23.09	1.34	1.22	26.65
Bi-Weekly	8.37	31.94	1.69	0.95	19.59
Tri-Weekly	7.30	42.9	2.01	0.87	24.56
Monthly	7.81	65.8	2.56	0.86	22.50
Panel B: MaxCor Portfolio for various frequencies					
	Ann. Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Daily	19.69	7.14	0.48	2.36	16.90
Bi-Daily	13.36	9.96	0.66	1.70	32.64
Weekly	13.32	24.84	1.09	1.62	28.48
Bi-Weekly	9.51	36.12	1.47	1.24	23.16
Tri-Weekly	8.86	50.64	1.80	1.15	16.95
Monthly	6.91	61.20	2.35	0.86	21.5
Panel C: AvgCor Portfolio for various frequencies					
	Ann. Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Daily	27.97	9.79	0.70	2.21	28.64
Bi-Daily	16.46	12.10	0.96	1.41	52.67
Weekly	16.41	30.19	1.56	1.38	37.26
Bi-Weekly	12.63	47.31	1.99	1.19	33.79
Tri-Weekly	11.65	65.80	2.39	1.13	24.91
Monthly	9.76	85.36	3.02	0.94	19.52

of lead-lag relationship erodes as the time between ranking of stocks and rebalancing the portfolios becomes longer. This result suggests that the market slowly absorbs the lead-lag relationship induced by the slow diffusion of information, and the residual lead-lag relationship at a monthly frequency becomes economically insignificant.

4.3.6.2 Slow information diffusion hypothesis and the speed of the market

Table 4.5 shows that the lead-lag relationship is strongest when we detect the lead-lag profiles and rebalance the portfolios both on daily frequencies. Portfolio performance weakens as the frequency of lead-lag detection and of portfolio rebalancing decreases. Thus, have daily lead-lag relationships consistently outperformed lead-lag relationships at lower frequencies throughout the entire historical period?

In this section, our study points to the time when the daily data-driven lead-lag portfolios first outperformed the lower-frequency portfolios. As Table 4.5 suggests, there is a point in time beyond which the daily lead-lag portfolios outperform all lower-frequency lead-lag portfolios; however, in the earlier years of our study, the daily lead-lag portfolio does not consistently outperform lower-frequency portfolios. For example, the daily Lévy-area portfolio begins to outperform the weekly Lévy-area portfolio in the mid to late-1970s; see Figure 4.5.

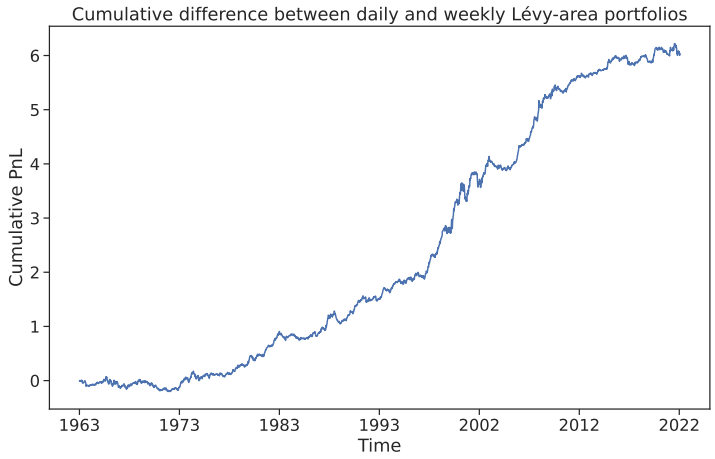


Figure 4.5: Cumulative sum of the difference between the daily lead-lag Lévy-area portfolio returns and the weekly lead-lag Lévy-area portfolio returns

To further explore how the frequency of lead-lag detection and rebalancing affects the returns of the three data-driven lead-lag portfolios, we record the first time these lead-lag portfolios outperform each of their corresponding lower-frequency portfolios. In particular, we track the initial occurrence of a cumulative excess return of 50% and 100% in the daily portfolio compared with the lower-frequency portfolios in Table 4.6.

Table 4.6 presents the first month after January 1963 when the cumulatively gained profit of daily lead-lag portfolios is at least 50% and 100% above the cumulative returns of each of the lower-frequency portfolios. We also record the first month when the daily lead-lag

Table 4.6: Dates when daily lead-lag portfolios first outperform lower frequency lead-lag portfolios†

Panel A: First occurrence of 50% cum. excess return between daily and lower-freq. portfolios						
	Raw	Bidiurnally	Weekly	Bi-weekly	Tri-weekly	Monthly
Lévy-area	1968.07	1973.04	1981.09	1980.09	1978.10	1971.02
Max Cross-Corr	1968.03	1974.07	1996.02	1982.07	1975.08	1970.08
Avg Cross-Corr	1968.02	1974.01	1995.06	1985.02	1982.08	1973.12
Panel B: First occurrence of 50% cum. excess return between daily and lower-freq. portfolios						
	Raw	Bidiurnally	Weekly	Bi-weekly	Tri-weekly	Monthly
Lévy-area	1973.04	1974.09	1987.10	1981.11	1981.03	1974.12
Max Cross-Corr	1971.04	1990.08	1999.08	1985.09	1983.02	1974.07
Avg Cross-Corr	1970.06	1982.12	1996.02	1987.10	1984.07	1981.08

† We use format YYYY.MM to express year and month.

portfolio cumulatively earns over 50% and 100% profit under the “Raw" column.

With the exception of bidiurnal lead-lag portfolios, we observe a clear trend in the time when the daily lead-lag portfolios start to outperform the lower-frequency portfolios. The cumulative returns of the daily portfolios outperform that of the monthly portfolios first, starting from the 1970s, and then outperform those of the tri-weekly, bi-weekly, and weekly portfolio in chronological order during the 1980s and the 1990s. For bidiurnal portfolios, however, we observe that the daily lead-lag portfolios outperform the bidiurnal portfolios much sooner than the daily portfolios outperform the weekly, and the tri-weekly portfolios.

Results in Table 4.6 lend further support to the slow information diffusion hypothesis, which asserts that lead-lag relationships are caused by the different speeds with which information is impounded into the leaders and the followers. In such cases, it takes between a week and a month for most of the information to diffuse to the followers in the early 1970s; over time, some information is impounded on the next day of trading whereas some residual information still takes weeks to trickle down to the followers. Over time, more information is absorbed and reflected in the prices of the followers on the next trading day and less relevant information is left to be captured at the lower frequencies, which causes the daily

rebalanced portfolio to outperform the lower frequency portfolios. The timings of changes in relative portfolio performances coincide with times when technological advancements enabled faster trading activities such as electronic execution and high-frequency trading, see McGowan (2010).

On the other hand, the bidiurnal frequency presents a unique dynamic. The literature does not compare performance of portfolios rebalanced at daily and bidiurnally, weekly, or monthly periods. Smith and Desormeau (2006) and Dennis et al. (1995) explore rebalancing periods over much slower frequencies such as annually or once every two years. We conjecture that if an investor possesses the technological capability to trade on a bidiurnal basis, it is likely that they can trade daily. As a result, market participants, especially informed traders who can trade at higher frequencies may find little incentive to trade bidiurnally because trading daily allows them to benefit from information asymmetries quicker than trading of lower frequencies. Consequently, the daily portfolios outperform the bidiurnal portfolio faster than they outperform the lower-frequency portfolios; we leave the discussion of this hypothesis to future work.

4.4 Robustness Analysis

In this section, we present the results of a number of robustness checks and alternative specifications in our study. The appendix includes further robustness analysis.

4.4.1 Execution cost and trading turnover

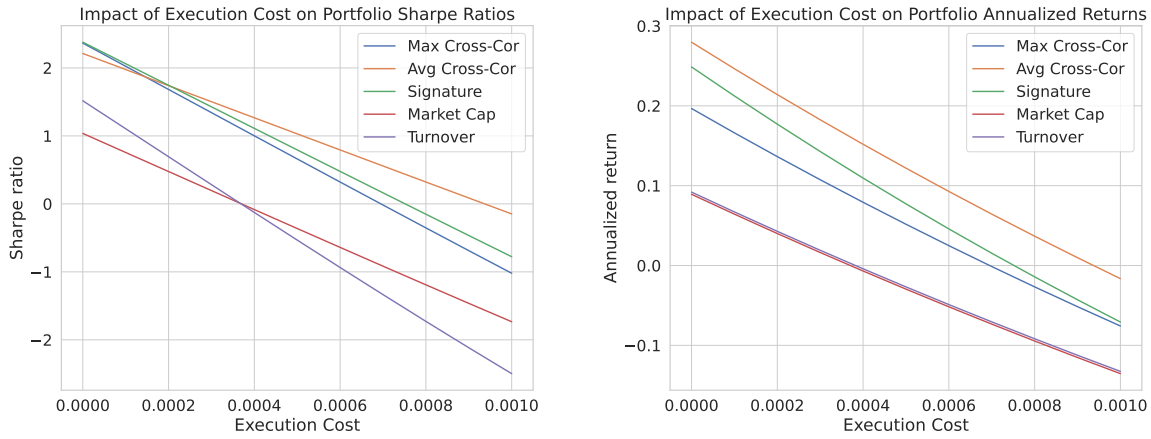
Here, we explore the impact of trading frictions on the performance of the lead-lag portfolios. We study performances of both the data-driven portfolios and the benchmark

portfolios as a function of trading frictions. To estimate the execution costs for a given portfolio, we first estimate the execution costs for trading one dollar worth of equity in the US market and then multiply this value by the turnover of the portfolio. The turnover of a portfolio that trades $d_1^t, \cdots, d_n^t$ dollar value in stocks $s_1, \cdots, s_n$ at time t is defined as

$$\text{TVR}_t = \frac{\sum_{i=1}^{n} |d_i^t - d_i^{t-1}|}{\sum_{i=1}^{n} |d_i^{t-1}|}, \tag{4.7}$$

where $|d_i^t - d_i^{t-1}|$ is the change in dollar value held by the portfolio on asset s_i in between rebalances.

For each portfolio, we compute its daily turnover and explore the portfolio's performance as a function of per-dollar execution cost. Figure 4.6a shows Sharpe ratios of the portfolios as a function of execution costs, and Figure 4.6b shows the returns of the portfolios as a function of execution costs. For each strategy, Table 4.7 presents the mean and standard deviation of its daily turnover ratio.



(a) (b)

Figure 4.6: (a) Sharpe ratio of portfolios as a function of per-dollar execution costs. Per-dollar execution cost is the cost incurred when performing a single trade on a portfolio of one dollar. (b) Annualized return of different portfolios as a function of per-dollar execution costs.

Table 4.7: Turnover ratio of various portfolios

	Avg. turnover ratio (%)	Avg. change in portfolio composition (%)	Sign flips (%)
Max Cross-Cor	102.5	22.6	44.3
Avg Cross-Cor	104.5	26.7	44.1
Lévy-area	111.3	23.6	50.4
Market Cap	95.4	1.21	44.2
Turnover	94.4	1.22	43.3

Figures 4.6a and 4.6b show that the data-driven lead-lag portfolios are significantly more robust to execution costs than the benchmark portfolios in terms of Sharpe ratio and annualized return. The Market Cap portfolio and the Turnover portfolio deliver losses when execution costs on each dollar traded is larger than 4 bps;[12] on the other hand, the data-driven portfolios are profitable when execution costs are 7 bps or lower, and their Sharpe ratio remains above 1 when execution costs are 4 bps.

We analyze the turnover ratio of the data-driven portfolios and of the benchmark portfolios in Table 4.7. We split the turnover ratio of the portfolios into two components, change in portfolio composition and sign flips. When the sign of the signal changes (i.e., the sign of the return of the leaders change), the turnover ratio of the portfolio is 2 regardless of the changes in the composition of the portfolio; on the other hand, when there is no sign flip, the turnover ratio of the portfolio is the proportion of the followers that change from the previous day's portfolio. Table 4.7 reports the average daily turnover ratio, average proportional change in portfolio composition, and the proportion of days when the sign of the signal flips.

Table 4.7 shows that ratio of turnover of the the data-driven portfolios is higher than that of the Market Cap portfolio and the Turnover portfolio. The composition of the data-driven lead-lag portfolios tends to change more than the baseline portfolios, but their proportion

[12]Currently, industry standard on trading a portfolio on large cap US equities is less than 3 bps per dollar. The fees for taking short positions is around 1 bps per dollar.

of sign flips is similar. Figures 4.6a and 4.6b show that the higher turnover ratio is not sufficient to offset the difference in performance between the data-driven portfolios and the baseline portfolios.

Finally, there is no bid-offer spread during the close auction, so our portfolios are not affected by the bid-offer spread that one would face during regular trading hours. Specifically, during the closing auction, buy and sell orders are collected and at a specified closing time, these orders are matched and executed at a single close price.

4.4.2 Exposure to illiquidity premium and Fama–French Factors

Amihud (2002) suggests that stock illiquidity positively affects stock excess return, and this phenomenon is referred to as illiquidity premium. In this section, we explore if the lead-lag portfolios we construct can be explained by the illiquidity premium of stocks and by other common risk factors.

We measure the illiquidity of stocks as the daily ratio of absolute stock return to its dollar volume, see Amihud (2002). Next, we construct a portfolio that consists of a long position in stocks with the highest 20% illiquidity measure and a short position in stocks with the lowest 20% illiquidity measure; we denote by $\mathrm{R}_t^{\mathrm{liq}}$ the returns of such portfolio for each day t.

We use linear models to explore if the illiquidity portfolio adds value in explaining the returns of our lead-lag portfolio, and we also explore if the returns of our lead-lag portfolio can be explained by the benchmark lead-lag portfolios discussed in section 4.3.2, and by

the Fama–French five factors. Consider the models

$$\mathrm{R}_t^{\text{Lévy}} = \alpha + \beta_{cap}\,\mathrm{R}_t^{\text{cap}} + \beta_{tvr}\,\mathrm{R}_t^{\text{tvr}} + \epsilon_t\,, \tag{M1}$$

$$\mathrm{R}_t^{\text{Lévy}} = \alpha + \beta_{liq}\,\mathrm{R}_t^{\text{liq}} + \beta_{cap}\,\mathrm{R}_t^{\text{cap}} + \beta_{tvr}\,\mathrm{R}_t^{\text{tvr}} + \epsilon_t\,, \tag{M2}$$

$$\mathrm{R}_t^{\text{Lévy}} = \alpha + \beta_{liq}\,\mathrm{R}_t^{\text{liq}} + \beta_{cap}\,\mathrm{R}_t^{\text{cap}} + \beta_{tvr}\,\mathrm{R}_t^{\text{tvr}} + \sum_{\text{Fama5}} \beta_{FF5}\,\mathrm{R}_t^{\text{FF5}} + \epsilon_t\,, \tag{M3}$$

where $\mathrm{R}_t^{\text{Lévy}}$ is the return of the Lévy-area lead-lag portfolio on a daily rebalancing frequency, $\mathrm{R}_t^{\text{cap}}$ and $\mathrm{R}_t^{\text{tvr}}$ are returns of the two benchmark lead-lag portfolios based on market capitalization and stock turnover ratios, and $\mathrm{R}_t^{\text{FF5}}$ are the returns of the five-factor Fama–French portfolios.

Table 4.8: Regression of Lévy-area lead-lag portfolio against various benchmarks

	Intercept [bps]	β_{liq}	β_{cap}	β_{tvr}	β_{Mkt}	β_{SMB}	β_{HML}	β_{RMW}	β_{CMA}	R^2 [%]
M1	$10.1^{\ddagger}$	-	$0.401^{\ddagger}$	$-0.026^{\dagger}$	-	-	-	-	-	12.3
	(22.26)		(45.08)	(-1.99)						
M2	$11.0^{\ddagger}$	$-1.495^{\ddagger}$	$0.397^{\ddagger}$	-0.016	-	-	-	-	-	12.5
	(22.8)	(-5.38)	(44.48)	(-1.24)						
M3	$11.0^{\ddagger}$	$-1.462^{\ddagger}$	$0.402^{\ddagger}$	-0.022^{*}	$0.0001^{\dagger}$	-0.00008	-0.0001	-0.0002	-0.0002	12.6
	(22.8)	(-5.25)	(44.71)	(-1.70)	(2.56)	(-0.85)	(-1.05)	(-1.77)	(-1.41)	

Table 4.8 reports results from regressions of daily returns for M1, M2, and M3. Data are from CRSP, and returns are measured from close-to-close. The t-statistics are reported in parentheses, and statistical significance at the 1%, 5%, and 10% level is indicated by ‡, †, and *, respectively.

There is little difference in the intercept and the R^2 across the three models. Overall, we observe that on average, around 10 to 11 basis points of the daily returns of the Lévy-area lead-lag portfolio cannot be explained by the risk factors and benchmark portfolios we include in the models. Notably, the Lévy-area lead-lag portfolio exhibits a positive exposure to the market capitalization lead-lag portfolio, indicating a propensity to select smaller-

sized stocks. On the other hand, it demonstrates a negative exposure to the Turnover ratio lead-lag portfolio and the illiquidity premium portfolio, suggesting a preference for more liquid stocks and for stocks with higher turnover ratios.

Results for M2 and M3 show that the returns of the Lévy-area lead-lag portfolio are not statistically significantly explained by the Fama–French five factors after considering lead-lag relationship related to size and liquidity. We observe no significant difference in the intercept and R^2 between the two models, and the coefficients on the Fama–French five factors are both small in magnitude and statistically insignificant.

Furthermore, comparisons between M1 and M2 imply some collinearity between the illiquidity premium portfolio and the Turnover ratio lead-lag portfolio. The addition of the illiquidity premium portfolio to the regressions results in a 0.1% increase in R^2, and the coefficient on the Turnover ratio lead-lag portfolio becomes statistically insignificant. These findings suggest that the illiquidity premium portfolio does not contribute significantly to explaining the returns of the Lévy-area lead-lag portfolio beyond the information provided by the two benchmark lead-lag portfolios included in the analysis.

4.4.3 Alternative selection of sets of stocks

In this subsection, we show that the lead-lag relationship we detect is robust to alternative selection of the set of stocks. In our data-driven baseline results, the set of stocks in a given trading day consisted of the top 25 percentile in terms of market capitalization in that day. This choice ensures that the portfolios we construct can be traded at scale in the market.

To assess the robustness of our results for a different set of stocks, we select those that rank

at the top 25 percentile in terms of both volume traded and turnover ratio. We obtain a smaller set of stocks than the one used above because volume and turnover ratio vary more than market capitalization. On average, this new set of stocks contains approximately one third of the stocks that we use in the main results above. Table 4.9 reports the results for this alternative set of stocks.

Table 4.9: Performance of lead-lag portfolios in the alternative set of stocks

Panel A: Global Lead-Lag Portfolios					
	Ann. Return (%)	Return (bps/day)	Volatility (%)	Sharpe	Max Drawdown (%)
Max Cross-Cor	35.57	12.08	1.90	1.01	82.47
Avg Cross-Cor	48.69	15.75	2.09	1.20	76.44
Lévy-area	51.85	16.59	2.08	1.27	91.31
Market Cap	4.62	1.80	1.77	0.16	99.60
Turnover	21.48	7.72	1.23	1.00	52.49
Panel B: Clustered Lead-lag Portfolios					
	Ann. Return (%)	Return (bps/day)	Volatility (%)	Sharpe	Max Drawdown (%)
Max Cross-Cor	26.03	9.18	1.08	1.35	46.28
Avg Cross-Cor	27.61	9.68	1.16	1.33	36.10
Lévy-area	31.45	10.86	1.18	1.46	40.53
Industry	4.55	1.77	1.39	0.21	87.55

Panel A of Table 4.9 compares the performance of the lead-lag portfolios without clustering. The daily returns for the portfolios range from 1.80 bps per day for the Market Cap portfolio to over 17 bps per day for the Lévy-area portfolio. The data-driven portfolios outperform the two benchmark portfolios for all criteria except for volatility and maximum drawdown. The performances of portfolios with the alternative set of stocks show the same relative hierarchy as in the main results presented above. In particular, the Lévy-area portfolio achieves 51.85% annualized return. The Sharpe ratio of the data-driven portfolios is higher than those of the two portfolios based on firm characteristics. This shows that the data-driven methods capture economically significant lead-lag relationships in the alternative set of stocks. In particular, the Sharpe ratio and annualized returns of the Lévy-area portfolio are the highest among the three data-driven methods

in panel A of Table 4.9.

Panel B of Table 4.1 compares the performances of various clustered lead-lag portfolios. The performances of the Max Cross-Cor, Avg Cross-Cor, and Lévy-area portfolios are economically significant, with returns ranging from 9.18 to 10.86 bps/day. However, the Industry portfolio is not economically significant, with a return of less then 2 bps per day. The data-driven portfolios, Max Cross-Cor, Avg Cross-Cor, and Lévy-area, outperform the Industry portfolio in all categories. In particular, the compounded return of the Lévy-area portfolio is 31.45%, and its Sharpe ratio of 1.46 is the highest. This suggests that the data-driven methods used in constructing the lead-lag portfolios capture economically significant lead-lag relationships, with the Lévy-area method being particularly effective, and are robust to this alternative set of stocks.

When compared with the results in Table 4.1, the performance of the portfolios for the alternative set of stocks is less economically significant. Although the annualized returns of portfolios in panel A of Table 4.9 are almost twice the value of those in Table 4.9, the volatility of returns of portfolios for the alternative set of stocks is almost four times the volatility of the portfolios in Table 4.1. Therefore, the Sharpe ratio results in Table 4.9 are half the value of those shown in Table 4.1. This result shows that including more stocks benefits the performance of lead-lag portfolios, while the gap in performance between the data-driven portfolios and the portfolios based on firm characteristics remains robust to the selection of the set of stocks.

Furthermore, the benchmark portfolios with the alternative set of stocks perform worse than those with the set of stocks discussed above; meanwhile, the data-driven portfolios still remain economically significant. This shows that the data-driven methods in lead-lag

detection are robust and transferable to various sets of assets.

4.4.4 Intermediate lead-lag relationships

In this section, we show that the lead-lag relationship we find is robust to considering the so-called “intermediate lead-lag relationships".

Recall that when identifying leaders and followers among the assets, we apply ranking methods to sort the assets from most likely to lead to most likely to follow and construct a portfolio that uses the returns of assets that rank at the top to predict the returns of the assets that rank at the bottom. While this method provides a criterion to evaluate lead-lag relationships, it neglects assets that rank in the middle of the sorting. Consider, for example, a set of three assets a, b, c, where asset b ranks in the middle, follows a, and leads c. In this case, asset b exhibits an “intermediate lead-lag relationship" as it is not included in the lead-lag portfolio, but possesses lead-lag relationships with other assets. It is also possible that asset b has no relationship with a or c.

Here, we extend the lead-lag detection mechanisms we designed above to study intermediate lead-lag relationships. For each trading day, we rank the assets into quantiles, denoted by $Q1, Q2, Q3, Q4$, respectively. Instead of using the previous return of $Q1$ to predict the return of $Q4$ as in Algorithm 1, we construct three portfolios in which we sequentially use $Q1$ to predict $Q2$, $Q2$ to predict $Q3$, and $Q3$ to predict $Q4$. We finance each of the three portfolios with a position in the SPY ETF and take the average of the three portfolios to construct the final portfolio, see Figure 4.7.

The portfolio illustrated in Figure 4.7 considers the intermediate lead-lag relationships, so measuring the performance of the portfolio can help to understand if these relationships

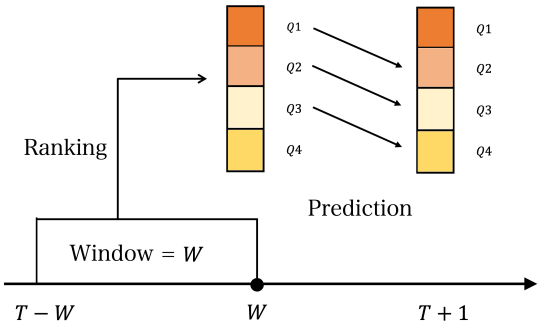


Figure 4.7: Portfolio construction of lead-lag portfolio with intermediate lead-lag relationships

are economically significant. Table 4.10 reports the performance of the Lévy-area, Max Cross-Cor, and Avg Cross-Cor, respectively.

Table 4.10: Intermediate Lead-lag Relationships

	Ann. Return (%)	Return (bps/day)	Vol. (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	16.1	5.91	0.40	2.37	15.8
Avg Cross-Cor	14.8	5.48	0.35	2.52	14.1
Lévy-area	20.4	7.36	0.40	2.89	14.7

The table shows that the lead-lag portfolios are economically significant after considering intermediate lead-lag relationships. However, compared with the results without considering intermediate lead-lag relationships in Table 4.1 where the top 20% of stocks are considered leaders and the bottom 20% of stocks are considered followers, the portfolios in Table 4.10 show lower return, but obtain lower volatility, higher Sharpe ratio, and lower maximum drawdown. Thus, the results show that our method is robust to considering intermediate lead-lag relationships and that intermediate lead-lag relationships do exist in the equities market.

4.5 Conclusion

This paper presented a method to detect linear and nonlinear lead-lag relationships in the US equity market. In contrast to the extant literature, which uses firm characteristics such

as market capitalization, trading turnover, and trading volume to select leaders and followers, our method employs the Lévy-area between pairs of stock returns to infer which one in the pair is more likely the leader, and to quantify the strength of this relationship. We constructed a portfolio that uses the previous returns of the leaders to determine positions on the followers; and showed that they generate economically significant performances that outperform all benchmarks in the literature. The performance of our portfolios is robust to various alternative specifications in algorithms, hyperparameters, and data sets.

The performance of the lead-lag portfolios we construct cannot be fully explained by lead-lag effects generated by market capitalization, trading volume, or intra-industry; there is little overlap between the composition of our lead-lag portfolios and that of the benchmarks. Our results show that accounting for nonlinearities is key to determine the lead-lag relationships.

The lead-lag relationships we find change over time. The leader-follower identity of stocks in various sectors changes several times between 1963 and 2022. This finding further supports the necessity of data-driven lead-lag detection methods that capture dynamically evolving lead-lag relationships.

Finally, we examined the performance of our portfolios across various rebalancing frequencies, and the results provided empirical support to confirm the slow information diffusion hypothesis. Specifically, the performance of portfolios decreases as both the ranking and the rebalancing are performed less frequently.

Our research leads to various future directions of work. One, explore lead-lag relationship on intra-day data, for alternative markets, and in alternative asset classes. Such a study might explain the relative differences in performance of lead-lag portfolios that are rebal-

anced at various frequencies over time. Two, understand the analytical properties of the nonlinear relationship between leaders and followers with methods such as smoothing or other non-parametric approaches. This direction can help to understand the source and composition of lead-lag relationships from the perspective of lagged factor exposures or other asset pricing models. Finally, we intentionally kept the construction of portfolios simple in this paper; future work is to study advanced portfolio construction methods that optimize the portfolio positions based on the lead-lag relationships we uncover.

5 | Volume Shocks and Overnight Returns

5.1 Introduction

Understanding the relationship between trading volume and stock returns provides insights on market dynamics, investor behavior, and the flow of information. A particular aspect of this relationship, known as *abnormal volume* or *volume shocks*, refers to the surprise in trading volume that deviates from its expected levels. A strand of the literature suggests that shocks in volume reflect heightened investor attention and information asymmetry, thereby influencing asset prices (Blume et al., 1994; Gervais et al., 2001; Kaniel et al., 2012; Kothari et al., 2016).

Recent developments in the literature revisit asset pricing features during the overnight period—a time when markets are closed but news and information continue to accumulate.[1] Overnight returns (stock returns from close to open) exhibit distinct patterns from those of intraday returns, and the difference is often driven by factors such as information disclosure, liquidity constraints, and investor sentiment (Branch and Ma, 2006; Cooper et al., 2008; Berkman et al., 2012; Lou et al., 2019; Boyarchenko et al., 2023). While this strand of the literature studies many asset pricing phenomena, the relationship between volume and overnight stock returns remains unclear. This raises questions about how volume shocks impact stock returns when the market is open versus when the market is closed.

For each stock, empirical studies show that periods of unexpected high trading volume are

[1] For example, most earnings announcements happen during the overnight session, see Doyle and Magilke (2008).

often followed by periods of high returns and vice versa. For example, Gervais et al. (2001) attribute this to increased visibility, which drives greater demand and higher expected returns. Similarly, Kaniel et al. (2012) document this phenomenon across international markets, linking volume shocks to improved investor recognition. However, we show that while a positive correlation exists between volume shocks and subsequent stock returns during the overnight period, this relationship does not hold intraday.

Our finding challenges the explanation that the link between volume shocks and stock returns is due to investor attention and its effect on the firm's cost of capital. Barber and Odean (2008), Fang and Peress (2009), and Israeli et al. (2022) argue that heightened attention expands a firm's investor base, which lowers its cost of capital and increases realized returns.[2] However, the difference in the link between volume shocks and stock returns during overnight and intraday sessions suggests more complex dynamics. Our findings imply that the traditional explanation may be incomplete, requiring a reassessment of how volume shocks influence asset pricing.

We define volume shocks as the relative deviation of daily trading volume from its exponential moving average (EMA)

$$\text{Volume Shock}_t = \frac{\text{Volume}_t}{\text{EMA Volume}_{t-1}} - 1\,. \tag{5.1}$$

This metric captures unexpected surges and decreases in trading activity, which the literature links to various market characteristics such as investors' attention.

We use volume shocks data to analyze the performance of portfolios during the overnight and intraday periods, and we uncover a robust pattern: stocks experiencing large (small)

[2]See Merton (1987) for the initial discussion on investor attention and cost of capital.

volume shocks during trading hours generate significantly high (low) returns over the subsequent overnight session, but this effect does not manifest during the next intraday trading hours.

Our results have several implications for the existing literature. First, they challenge the notion that the relationship between volume shocks and subsequent stock returns is primarily a consequence of an increase in investor attention that lowers the cost of capital. If this were the case, one would expect the positive correlation between volume shocks and stock returns to manifest during regular trading hours on the next trading day when investors can act on new information. The absence of significant next day intraday returns associated with volume shocks suggests that other factors are at play.

Second, our findings contribute to the growing body of research examining the unique characteristics of overnight returns. Prior studies highlight that overnight returns often reflect information assimilation processes that differ from those during the trading day. Branch and Ma (2006) document that the overnight return is an anomaly in itself, exhibiting patterns not explained by traditional models. Hendershott et al. (2020) revisit the Capital Asset Pricing Model (CAPM) and highlight significant differences in the dynamics of asset prices between night and day, and they show that the market risk premium is positive overnight and negative intraday. Our work contributes to this strand of the literature; we show that volume shocks—typically considered within the context of trading hours—have a pronounced effect on returns precisely when markets are closed.

Third, our study offers insights into market efficiency and the role of trading constraints. Market friction is one explanation for the difference in the dynamics of intraday and overnight stock returns. Market frictions such as liquidity constraints, the presence of noise

traders, and institutional trading practices that delay the full incorporation of information into prices until the start of the next trading session (Hong and Stein, 1999; Hirshleifer et al., 2011; Bogousslavsky, 2021). For example, Barardehi et al. (2022) suggest that momentum and reversal patterns are driven by different day and night signals, which supports the idea that stock returns during overnight and intraday sessions may be driven by different activities and market dynamics.

Our research methodology involves comprehensive empirical analyses, including portfolio construction based on volume shocks, regression analyses against well-known risk factors, and predictive models of volume that address non-tradability issues. In particular, we build a volume shock prediction model to address that one cannot have perfect foresight of volume shock at the closing auction when one builds the overnight portfolios.

We sort stocks by their volume shocks to form equally-weighted and value-weighted portfolios. Next, to test our claims, we regress portfolio returns against well-known risk factors—such as market, momentum, reversal, and size—and we find that traditional asset pricing models, including those in Fama and French (2023), do not explain the positive overnight returns associated with volume shocks. We also study subsamples of stocks by their size and show that the relationship between volume shock and subsequent overnight returns is robust regardless of the market capitalization of stocks.

In addition to considering portfolios, we employ Fama–MacBeth and pooled regression analyses to explore the average expected stock return associated with volume shocks. The results consistently show a significant and positive overnight risk premia associated with volume shocks, reinforcing the notion that exposure to volume shocks provides distinct insight to stock returns between trading and non-trading hours. Additionally, while the

volume-shock-sorted portfolios report near-zero intraday profit and loss (PnL), Fama–MacBeth and pooled regression report statistically significantly negative average stock excess return associated with volume shock when the market is open.

In addition to conventional asset pricing exercises, we investigate if including the return of volume-shock-sorted portfolios as one explanatory variable improves explainability on the returns of the market portfolio. Including volume shocks significantly increases the model's explanatory power (R^2) for overnight returns but not for intraday returns. Our approach is similar to that of Savor and Wilson (2014), who show that asset prices behave differently on days when macroeconomic news is scheduled for announcement.

The portfolios we use to study relationships between volume shocks and stock returns are non-tradable because volume shocks are realized after the closing auction, and we mark our trades from the closing auction to the opening auction. To address this limitation, we employ predictive models of volume shock. For each day, we use information up until the open auction of the day to predict the volume shock of that day, and then we use the predicted volume shock to decide how much to buy of each stock. In that way, one has the whole intraday period for trade execution to minimize trading costs and price impact, and we ensure that there is no lookahead bias in our process. While we do not seek to create the best volume shock prediction model, we show that the positive correlation between volume shocks and subsequent stock overnight returns is not a result of the inability to trade stocks at the close after the market learns the trading volume of the day.

We use linear regression, LightGBM (Ke et al., 2017), and TabNet (Arik and Pfister, 2021) to forecast volume shocks and use these predictions to construct implementable trading strategies. With a linear model, one can build portfolios that achieve over 90% of the

performance (in terms of Sharpe ratio) of the portfolio that trades stocks every day at the close auction with perfect foresight of volume shocks. Moreover, volume prediction models with higher predictive accuracy improve portfolio performance, and the relationship between predictability of volume and portfolio performance is concave. The R^2 of advanced models like TabNet is twice that of linear models, and the Sharpe ratio obtained with TabNet is marginally higher than that obtained with the linear model, which suggests diminishing returns from increased predictive precision of volume shock.

Our study aligns with and extends several strands of the finance literature. Previous research explores the role of volume shocks in price discovery and in return predictability. For instance, Gervais et al. (2001) find that unusually high trading volumes are followed by significant, positive future stock returns, indicating that volume shocks may signal persistent information that the market gradually incorporates. Lerman et al. (2008) link the high-volume return premium to post-earnings announcement drift, suggesting that volume shocks capture delayed reactions to fundamental information.

In the context of investor attention, Barber and Odean (2008) report that individual investors are net buyers of attention-grabbing stocks, leading to predictable patterns in stock returns. Fang and Peress (2009) show that media coverage influences investor attention, which affects stock returns and the firm's cost of capital. These studies imply that volume shocks, as proxies for investor attention, should have a consistent effect on returns during trading hours.

Our findings challenge this perspective. We show that the positive correlation between volume shocks and stock returns is confined to the overnight period. This discrepancy in findings suggests an incompleteness in the traditional interpretation that investor at-

tention, proxied by volume shocks, affects stock returns by reducing the firm's cost of capital. It raises the possibility that other mechanisms—such as delayed information processing or overnight risk premia—are responsible for the observed patterns. Boudoukh et al. (2019) support this notion. They show that firm-specific news significantly impacts overnight volatility of returns, indicating that information flow during non-trading hours is potentially another crucial factor that explains overnight stock returns. Therefore, one possibility is that volume shock is a proxy for people's opinions on events that might happen in the following overnight hours.

Recent literature focuses on the overnight period to understand asset pricing anomalies. Berkman et al. (2012) examine overnight returns and the hidden cost of buying at the opening auction, highlighting that significant price adjustments often occur outside trading hours due to the behaviour of retail investors. Lou et al. (2019) and Akbas et al. (2022) investigate the tug of war between overnight and intraday expected returns. They find that overnight returns play a crucial role in momentum profits and future stock returns. Our study provides evidence that volume shocks have a unique and significant effect on overnight returns, further emphasizing the importance of the non-trading period in asset pricing.

Additionally, our research intersects with studies on market microstructure and information asymmetry. Hong and Stein (1999) develop a model of underreaction, momentum trading, and overreaction in asset markets, suggesting that gradual information diffusion can lead to delayed price adjustments. Hirshleifer et al. (2011) explore limited investor attention and stock market misreactions, and they propose that constraints on attention can delay responses to information. These theories support the idea that the market's reaction to volume shocks may be postponed until the overnight period due to cognitive

or institutional factors.

Our study also relates to the literature on liquidity and trading constraints. Amihud and Mendelson (1986) show that expected stock returns are inversely related to liquidity, with less liquid stocks requiring higher returns to compensate investors for trading costs. During the overnight period, liquidity is inherently limited, and trading constraints are more pronounced. This environment may exacerbate the impact of volume shocks on returns because price adjustments occur when the market reopens, reflecting accumulated information and liquidity considerations.

Moreover, Boyarchenko et al. (2023) document overnight drift, where the mean returns of U.S. equities are large and positive during the opening hours of European markets. They attribute this to inventory risk models and asymmetric reactions to demand shocks. Our findings complement their result. We show that volume shocks could be one of the factors contributing to the overnight drift.

In summary, our research provides new insights into the complex relationship between volume shocks and stock returns. We challenge the theories that the positive correlation between volume shocks and stock returns is attributed to investor attention during trading hours; our results show that this phenomenon is only statistically and economically significant overnight. Our findings suggest that alternative mechanisms—such as delayed information processing, overnight risk premia, or liquidity constraints—may better explain the observed patterns between volume shock and future stock returns.

The remainder of the paper is organized as follows. Section II describes the data we use in our study. Section III describes our methodology, including the construction of volume shock measures and portfolio formation techniques. Section IV presents the empirical

results, discussing the implications of our findings. Section V concludes the paper and suggests future research.

5.2 Data

Data are from the Center for Research in Security Prices (CRSP) between January 2000 and December 2022. We include all common share stocks (share codes 10 and 11) traded in NYSE, Amex, and NASDAQ. For each stock, we use the daily: low price "BIDLO", high price "ASKHI", close price "PRC", open price "OPENPRC", trading volume "VOL", dividend-adjusted close-to-close return "RET", shares outstanding "SHROUT", and cumulative factor to adjust price "CFACPR". For each stock, the adjusted price is the quotient between the close price and the CFACPR, and the market capitalization of each stock is the product between their close price and the number of shares outstanding.

On day t, the overnight return for stock i is

$$\mathrm{R}_{i,t}^{\text{overnight}} = \frac{1 + \mathrm{R}_{i,t}^{\text{close-to-close}}}{1 + \mathrm{R}_{i,t+1}^{\text{open-to-close}}} - 1\,, \tag{5.2}$$

see Akbas et al. (2022) and Hendershott et al. (2020). Here, the intraday return of stock i on day $t+1$ is

$$\mathrm{R}_{i,t}^{\text{open-to-close}} = \frac{\mathrm{P}_{i,t}^{\text{close}} - \mathrm{P}_{i,t}^{\text{open}}}{\mathrm{P}_{i,t}^{\text{open}}}\,, \tag{5.3}$$

where $\mathrm{P}_{i,t}^{\text{close}}$ and $\mathrm{P}_{i,t}^{\text{open}}$ are the close and open prices on day t for stock i, respectively. The close-to-close return $\mathrm{R}_{i,t}^{\text{close-to-close}}$ is the dividend and other corporate-actions-adjusted return provided by CRSP.

When the open price of a stock is missing, we use the previous day's close price of the

stock, in which case the overnight return from the previous day is zero and the intraday return of the stock is the dividend-adjusted close-to-close return.

Next, the volume shocks for each stock are the deviation of the current day's volume during trading hours $V_{i,t}$ from the exponential moving average (EMA) of the stock's previous volumes. The EMA, which assigns more weight weight to recent data, is given by

$$\mathrm{EMA}_{i,t} = \mathrm{EMA}_{i,t-1} + \left(1 - \exp\left(-\frac{\ln 2}{60}\right)\right) \times \left(V_{i,t} - \mathrm{EMA}_{i,t-1}\right),$$

where $\mathrm{EMA}_{i,t-1}$ is the previous day's EMA of volume, and as time evolves, its weight decreases exponentially with a half-life of 60 trading days. The volume shock $S_{i,t}$ for stock i on day t is

$$S_{i,t} = \frac{V_{i,t}}{\mathrm{EMA}_{i,t-1}} - 1\,.$$

The volume shock $S_{i,t}$ represents the "surprise" of the realized volume for stock i on day t compared with an estimation of its "normal" trading volume given by the EMA of the stock up to day $t-1$ (inclusive).

5.3 Methodology

We construct ten single-sorted portfolios and ten double-sorted portfolios to study how volume shocks on day t affect the overnight returns of stocks and their intraday returns on day $t+1$. We test the null hypothesis that volume shocks do not contain directional information about future stock prices. The alternative hypothesis is that positive (negative) volume shocks predict high (low) overnight and intraday returns in the future. To test the null hypothesis, we construct portfolios where stocks are single sorted by volume

shock in the previous trading day, and we construct portfolios that are double sorted by market capitalization and by volume shocks to control for size of stocks.

Single-sorted portfolios. For each day t, we use volume shocks $S_{i,t}$ to rank stocks into quintiles from large to small. Next, for stocks in each volume shock quintile, we build equally-weighted and value-weighted long-only portfolios to realize the next-day overnight return $\mathrm{R}_{i,t+1}^{\text{overnight}}$ and the next-day intraday return $\mathrm{R}_{i,t+1}^{\text{day}}$. Specifically, to realize next-day overnight returns, one buys stocks in the close auction of day t and sell stocks at the opening auction of day $t+1$; similarly, to realize intraday returns, one buys stocks in the opening auction of day $t+1$ and sells stocks in the close auction of day $t+1$.[3]

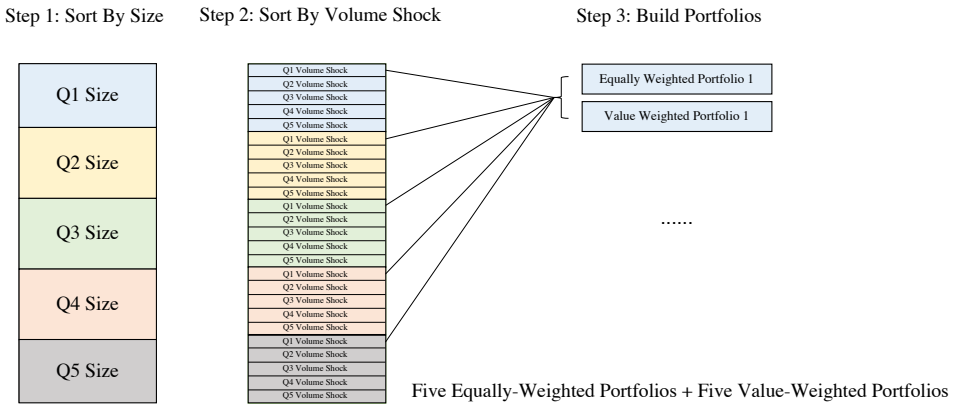


Figure 5.1: Three-step approach to construct double-sorted portfolios on size and volume shock.

Double-sorted portfolios. Next, we sort stocks into quintiles by size and by volume shocks to construct ten double-sorted portfolios. Blume et al. (1994) show that the properties of trading volume for large firms differ from those of smaller firms; therefore, one considers the effect of volume shocks on firms of different sizes separately. To this end, we construct ten double-sorted portfolios in three steps, see Figure 5.1. One, we use the market capitalization of stocks on day $t-1$ to rank all stocks on day t into quintiles, see

[3] Alternatively, this is equivalent to taking long positions in stocks and then separately measuring their overnight returns and intraday returns.

Gervais et al. (2001).

Two, within each market capitalization quintile, we further rank stocks from high to low into quintiles according to the volume shocks they experience on day t. Therefore, stocks are partitioned into a grid of twenty-five groups, each defined by a specific combination of size and volume shock quintile.

Three, one consolidate these twenty-five groups into five categories to obtain the double-sorted portfolios. Specifically, for each volume shock quintile, we gather all stocks in the quintile, regardless of their size quintile, to form five consolidated stock groups. Within each of these five groups, we construct equally-weighted and value-weighted portfolios. Thus, we build five equally-weighted portfolios and five value-weigthed portfolios where each portfolio contains a mix of stocks ranging from the largest to the smallest, according to their market size.

For overnight portfolios, we assume one can trade at the close of day t and unwind the position at the open of day $t+1$; and for intraday portfolios, we assume one can trade at the open of day $t+1$ and unwind the position at the closing auction of day $t+1$. We use the t-test to determine if the expected returns of each of these portfolios are statistically significant. For overnight portfolios, because volume shock is realized at the close auction, assuming one can build a portfolio at the close using volume shock is non-trivial; therefore, in the rest of the paper we refer to the portfolios we construct here as the "oracle" portfolios. The "oracle" portfolio assumes that one can trade at the close auction when the value of volume shock is not yet finalized. In practice, one either trades single-stock futures or makes predictions on volume shocks to trade the overnight portfolios. Below in Section 5.3.3, we provide one way of how to trade this portfolio in the market

where instead of assuming perfect knowledge of volume shock, we trade on predicted values of each stock's volume shock.

5.3.1 The effect of volume shocks on overnight and intraday returns

While we only report the performance of long-only portfolios, we build long-short portfolios to study if common risk factors explain the effect of volume shock on stock returns.[4] In particular, we assess if the market, size, long-term momentum, and short-term reversal factors explain the effect of volume shocks on future stock returns. For the single- and double-sorted portfolios, we take equally-weighted and value-weighted long positions on the first volume-shock quintile of stocks and short positions on the fifth volume-shock quintile stocks to construct long-short portfolios.

Next, we follow Fama and French (2023) to construct portfolios that represent the four risk factors we consider. The market factor portfolio uses all stocks in the data set to build value-weighted portfolios. We use single sortings to construct the size, momentum, and reversal portfolios. To construct the size portfolio, we first split the the stocks by their previous day's market capitalization into the bottom 30%, middle 40%, and top 30%, which we denote small, medium, and large stocks, respectively. Next, take equally-weighted long positions on the large capitalization stocks and take equally-weighted short positions on the small capitalization stocks to form the size portfolio. The long-term momentum and short-term reversal portfolios are constructed in the same way; we use the previous 12-day returns to sort the stocks from high to low and the previous one-day return to sort the stocks from low to high, respectively.

[4]Portfolio tradability is no longer a concern when focusing on explainability.

Now, we use the risk-factor-portfolios we construct to explain the returns on each long-short quintile portfolios constructed from volume shocks. Consider the model

$$\mathrm{R}_t^{\text{vs N/D}} = \alpha + \beta_{mkt}\,\mathrm{R}_t^{\text{mkt N/D}} + \beta_{size}\,\mathrm{R}_t^{\text{size N/D}} + \beta_{mom}\,\mathrm{R}_t^{\text{mom N/D}} + \beta_{rev}\,\mathrm{R}_t^{\text{rev N/D}} + \epsilon_t\,, \quad (5.4)$$

where $\mathrm{R}_t^{\text{vs N/D}}$ represents the overnight or the intraday return of the volume-shock long-short portfolio, $\mathrm{R}_t^{\text{mkt N/D}}$ is the market portfolio overnight or intraday, and $\mathrm{R}_t^{\text{size N/D}}$, $\mathrm{R}_t^{\text{mom N/D}}$, $\mathrm{R}_t^{\text{rev N/D}}$ are the overnight or intraday returns of the size-sorted, long-term momentum-sorted, and short-term reversal-sorted factor portfolios, respectively. We report the daily α, R^2 value, and the statistical significance of the coefficients to explore if these risk factors explain the returns of the long-short volume-shocks portfolios we construct above in the empirical results.

5.3.2 Relationship between volume shocks and the difference between overnight and intraday returns

We study if volume shocks can help explain the difference between overnight returns of stocks and their intraday returns. In particular, we are interested in two questions. One, can volume shocks explain the cross-sectional difference in the return of stocks intraday and overnight? In other words, are the signs and magnitudes of the risk premia of volume shock different when the market is open than when the market is closed? Two, can volume shocks explain the difference between intraday and overnight market returns? In other words, can the returns of the market be better explained with volume shock, conditioning on other risk factors? To answer the first question, we employ Fama–MacBeth regressions and pooled OLS, and to answer the second question, we use panel regressions.

With the Fama–MacBeth regression we study if volume shocks explain the cross-sectional difference between expected returns of individual stocks when the market is closed and when the market is open. Specifically, we use the Fama–MacBeth approach to estimate the beta of volume shock and the equity risk premium inferred from the volume shocks of stocks on their intraday and overnight returns. Consider the rolling-window time-series regressions on individual stock returns

$$\mathrm{R}^{\mathrm{N}}_{i,t+1} = \alpha^{\mathrm{N}}_{i} + \beta^{vs}_{i,t}\ \mathrm{S}_{i,t} + \beta^{\mathrm{Mkt}}_{i,t}\ \mathrm{R}^{\mathrm{Mkt\ N}}_{t+1} + \epsilon^{\mathrm{N}}_{i,t}\,, \tag{5.5}$$

where $\mathrm{R}^{\mathrm{N}}_{i,t+1}$ is the overnight return of stock i and $\mathrm{R}^{\mathrm{Mkt\ N}}_{i,t+1}$ is the overnight return of the market; using overnight returns to compute exposure to the market and to volume shock is consistent with the approach in Hendershott et al. (2020). Recall that $\mathrm{S}_{i,t}$ is the volume shock for stock i from the open auction of day t to the close auction of day t. The coefficients of the time series regression in (5.5) are re-computed every day with a lookback rolling window of one-year.

As the second step of the Fama–MacBeth approach, we use the estimated night beta to estimate the overnight and intraday risk premium of volume shocks. We use the estimated risk premium to assess if volume shocks affect the cross-sectional difference in the intra-day and overnight returns of stocks. Consider the cross-sectional regressions for the risk premium

$$\mathrm{R}^{\mathrm{N/D}}_{i,t+1} = \xi^{\mathrm{N/D}}_{0,t} + \xi^{\mathrm{N/D}}_{1,t}\ \hat{\beta}^{\mathrm{vs}}_{i,t} + \xi^{\mathrm{N/D}}_{2,t}\ \hat{\beta}^{\mathrm{Mkt}}_{i,t} + \epsilon^{\mathrm{N/D}}_{i,t}\,, \tag{5.6}$$

where $\hat{\beta}^{\mathrm{vs}}_{i,t}$ and $\hat{\beta}^{\mathrm{Mkt}}_{i,t}$ are the estimated volume shock beta and market beta from (5.5) on day t, and $\mathrm{R}^{\mathrm{N/D}}_{i,t+1}$ are the overnight and intraday returns of stocks, respectively.

In addition to the Fama–MacBeth regressions that are run separately on overnight and on

intraday returns, we estimate a single regression and directly test if the implied risk premia on volume shock when the market is closed and when the market is open are different. Specifically, we perform the regression

$$\mathrm{R}_{i,t+1}^{\mathrm{N/D}} = \xi_0 + \xi_1 \, \hat{\beta}_{i,t}^{\mathrm{vs}} + \xi_2 \, \hat{\beta}_{i,t}^{\mathrm{mkt}} + \xi_3 \, \mathrm{D}_{t+1} + \xi_4 \, \hat{\beta}_{i,t}^{\mathrm{vs}} \, \mathrm{D}_{t+1} + \epsilon_{i,t+1} \, , \tag{5.7}$$

where $\mathrm{R}_{i,t+1}^{\mathrm{N/D}}$ is either the overnight or the intraday return, $\epsilon_{i,t+1}$ is an error term, and D_{t+1} is an indicator variable equal to one for a intraday return and zero otherwise. This approach is consistent with that in Savor and Wilson (2014) and in Hendershott et al. (2020).

We record the t-statistic, R-square values, and estimate of the risk premium for the regressions from 2000 to 2022 and across various subsamples during this period. In particular, we observe how the risk premium of the market and volume-shock factors change over three subsamples from 2000 to 2009, from 2010 to 2019, and from 2020 onwards. The first subsample contains the market prior and during the financial crisis; the second subsample contains the market after the financial crisis and before Covid; and the third subsample contains the impact of Covid and the Russia-Ukraine war.

Rregressions (5.6) and (5.7) test for the relationship between the cross-sectional difference of stock returns and volume shock. Next, we explore if the difference between market level intraday and overnight returns is explained by volume shock. Consider the model

$$\mathrm{R}_t^{\mathrm{mkt\ N/D}} = \alpha + \beta_{vs} \, \mathrm{R}_t^{\mathrm{vs\ N/D}} + \beta_{mkt} \, \mathrm{R}_{t-1}^{\mathrm{mkt\ N/D}} + \beta_{size} \, \mathrm{R}_t^{\mathrm{size\ N/D}} + \beta_{mom} \, \mathrm{R}_t^{\mathrm{mom\ N/D}} + \beta_{rev} \, \mathrm{R}_t^{\mathrm{rev\ N/D}} + \epsilon_t \, . \tag{5.8}$$

We test the effect of the volume shocks portfolio $\mathrm{R}_t^{\mathrm{vs\ N/D}}$ on the fit of the model to explain

the returns of the market when it is closed and when it is open. Here, the volume-shocks portfolio is the long-short double-sorted value-weighted portfolio discussed in Section 5.3.1. If adding the return of the volume shocks portfolio $R_t^{\text{vs N/D}}$ in (5.8) significantly improves the explanatory power on the market for overnight sessions, but not for intraday sessions, then there is empirical support for the hypothesis that the observed difference in expected returns of the market between intraday and overnight sessions can be partially attributed to the difference in the exposure of stock returns to volume shocks when the market is open and when it is closed.

5.3.3 Machine learning model for volume shocks prediction and tradable portfolios

The portfolios we use in Section 5.3.1 to test if volume shocks affect the overnight returns of stocks cannot be traded. On day t, the volume shock of stock i is computed from open to close, and its overnight return is computed from the close of day t to the open of day $t+1$; thus, it is not feasible to trade the overnight returns of stocks from day t to $t+1$ because one cannot trade instantaneously at the end of the closing auction on day t. In practice, to trade the portfolios in Section 5.3.1, one needs to take positions in financial derivatives such as futures, which can be either infeasible or too costly.

In this section, we predict volume shocks to trade the portfolios in Section 5.3.1 in two steps. First, on day t, we use information up to the opening auction to predict the volume shocks of day t. Second, we use the predicted value of volume shocks to construct single-sorted portfolios in the same way as those in Section 5.3.1 (i.e., we sort the stocks by their predicted value of volume shocks into quintiles and build long-positions within each quintile). Thus, one can trade these portfolios anytime between the opening and the

closing auction of day t to realize the change in stock price overnight from the close of day t to the open of day $t+1$.[5]

The volume shock predictions for each stock in the first step are performed with statistical models ranging from linear regression to neural networks. Specifically, consider a set of features $\mathbf{F}_{i,t}$ for stock i on day t, which contains information of the stock up until the opening auction of day $t+1$. We use elements in $\mathbf{F}_{i,t}$ to predict the volume shock $S_{i,t+1}$ of the stock on day $t+1$. The predictive model considers the optimization problem [6]

$$\hat{f} = \underset{f \in \mathbb{H}}{\arg\min}\ \mathbb{E}_{i,t}\left[L(S_i,\ f(\mathbf{F}_{i,t}))\right] + \Omega(f)\,, \tag{5.9}$$

where L is a loss function, Ω is a regularization term, $\mathbb{H}$ is the space of functions from which one chooses the model f, and the expectation is over all stocks in the training sample. For example, in Ridge regression, $\mathbb{H}$ is the space of linear functions from the space of the vector of features to the space of real numbers, L is the sum of the squared errors between the predicted and realized values of volume shock, and Ω is the Euclidean norm of the coefficients of f.

We solve the optimization problem on the training sample to find the optimal f^*, which we use to predict the volume shocks of stocks on the test set and to evaluate the performance of the model.

Here, we split the data into training and testing sets. Data from January 2000 to December 2016 are for training and validation, and data after January 2017 are for out-of-sample

[5] In our experiments, we still mark returns from the closing auction to the opening auction. In practice, if one assumes intra-day returns have zero mean, one can trade at any time during the day to realize, in expectation, the same returns.

[6] Here we use the notation $\mathbf{F}_{i,t}$ to emphasize that we are carrying out a prediction task. In practice, we are only using information up to the open auction of t+1.

testing. Next, we construct a set of features with the dividend-adjusted price, close price, open price, daily high price, daily low price, daily volume, daily raw return, previous values of daily volume shock, and previous values of the exponential moving average of volume. Finally, we predict the volume shocks of stocks with these features as inputs to the machine learning models.

We group our features into five categories, where each category captures distinct aspects of market behavior and historical trends. The first category includes features related to intraday price movements, such as the absolute change in price from open to close and the spread between close and open prices. These metrics provide information about daily price volatility. The second category contains the momentum and volatility of variables, which incorporates rolling averages and standard deviation of the variables over various time windows. The third category contains lagged values of variables such as return and volume shocks, so our models capture temporal dependencies effectively. The fourth category contains the first difference and percentage change of the variables. The fifth category comprises features that capture other time series characteristics, including day-of-week, month-of-year, and moving average convergence divergence (MACD) of the variables. For the complete list of features, see Appendix A.8

We use five models to predict volume shocks for each individual stock. Our benchmark models are ordinary least squares and Lasso regression. The other machine learning models are LightGBM (Ke et al. (2017)), Multilayer Perceptron (MLP), and TabNet (Arik and Pfister (2021)). For an introduction to the machine learning models, see Appendices A.9 and A.10. We use grid-search cross-validation to select the appropriate hyper-parameters of our models.

After obtaining predictions of the volume shock for each stock on day t, we build a portfolio that buys stocks at the close of day t and unwinds the position at the open of day $t+1$.

Specifically, we sort stocks into quintiles by the predicted value of their volume shocks on day t, and then we take equally-weighted and value-weighted long positions on the first quintile of stocks. Recall that we can trade the stocks at the close auction of day t and close the position on the opening auction of day $t+1$ because we sorted stocks at the opening auction of day t; therefore, we circumvent the infeasibility in trading the volume-shock-sorted portfolios proposed in Section 5.3.1.

We trade a long-only portfolio instead of long-short portfolios as commonly proposed in the asset pricing literature. If our prediction of volume shocks is accurate, the trading volume of stocks in the lowest quintile by predicted volume shocks would be much lower than their normal volumes, so, potentially, trading these stocks is difficult due to limited liquidity. The market impact cost of trading iliquid stocks is high and could erode the profitability of the portfolio. On the other hand, the volume of the stocks in the first quintile is much larger than usual, so, everything else being equal, one can buy these stocks with little adverse price impact.

After constructing the portfolios, we compare their performance with that of the oracle portfolios we constructed in Section 5.3.1. Also, we assess if more advanced machine learning models predict volume shocks more accurately and if better predictions of volume shocks are positively correlated to less information loss in portfolio construction as compared to the "oracle" portfolios.

5.4 Results

5.4.1 High-volume premium on overnight and intraday returns

Table 5.1 shows the results for the single-sorted and double-sorted quintile oracle portfolios of Section 5.3.1. The table reports the annualized compound return, daily return, t-statistics of daily return, daily volatility, the Sharpe ratio of each of the quintile portfolios, and the performance of the market when the market is closed and when the market is open. We also report the overall return of the market overnight and during the day, which we refer to as "All Night" and "All Day" portfolios, respectively.

Panel A of Table 5.1 reports the performance of the volume-shock single-sorted oracle portfolio performances during the day and overnight for equally-weighted and value-weighted portfolios. For overnight returns, we observe a monotonic trend where the returns and Sharpe ratio of portfolios with stocks that undergo larger volume shocks tend to be larger than those with stocks that undergo smaller volume shocks. This finding applies to both equally-weighted and value-weighted portfolios. For equally-weighted portfolios, the annualized returns of stocks with volume shocks in the top 20% are over 70%, on average, as opposed to the stocks in the bottom 20% whose returns are negative. On the other hand, for value-weighted portfolios, the annualized returns of the portfolio in the top 20% of stocks by volume shock are 17%, while the annualized returns of the bottom quintile are negative. The returns of the top 20% volume-shock-quintile portfolio are also statistically significant with t-statistics of 22.82 and 5.74 for equally-weighted and value-weighted portfolios, respectively; and their Sharpe ratios are 4.76 and 1.20, respectively. The top 20% and top 40% volume-shock-quintile portfolios outperform the return of the market

Table 5.1: Performance of sorted volume shocks oracle portfolios

Performance of the quintile portfolios single- and double-sorted by volume shock. The night portfolios are measured daily from close to open, and the day portfolios are measured daily from open to close. Panel A reports results from single-sorted portfolios and Panel B reports results from double-sorted portfolios. Data are from CRSP.

Panel A: single-sorted oracle quintile portfolios

	Equally Weighted					Value Weighted				
	Comp. Ret.[%]	Ret. [bps/-day]	t-stat	Std. Dev.[%]	Sharpe	Comp. Ret.[%]	Ret. [bps/-day]	t-stat	Std. Dev.[%]	Sharpe
All Night	21.97	7.36	8.88	0.63	2.00	7.88	3.01	3.32	0.70	0.68
Night Q1	71.24	21.37	22.82	0.71	4.76	17.00	6.23	5.74	0.83	1.20
Night Q2	20.78	7.50	8.15	0.70	1.70	8.63	3.28	3.58	0.70	0.75
Night Q3	14.74	5.46	6.06	0.69	1.26	4.56	1.77	2.01	0.67	0.42
Night Q4	11.05	4.16	4.86	0.65	1.01	3.84	1.50	1.70	0.67	0.35
Night Q5	−4.39	−1.78	−2.79	0.49	−0.58	−0.19	−0.08	−0.10	0.61	−0.02
All Day	−2.83	−1.13	−0.85	1.39	−0.17	1.48	0.58	0.45	1.02	0.09
Day Q1	−4.95	−2.02	−1.39	1.10	−0.29	−2.47	−0.99	−0.66	1.14	−0.14
Day Q2	4.03	1.57	1.03	1.16	0.21	0.97	0.38	0.29	1.01	0.06
Day Q3	3.66	0.14	0.09	1.18	0.02	1.60	0.63	0.47	1.02	0.98
Day Q4	−5.08	−2.07	−1.35	1.17	−0.28	0.49	0.19	0.14	1.09	0.03
Day Q5	−0.60	−0.24	−0.22	0.83	−0.05	−6.11	−2.50	−1.85	1.03	−0.39

Panel B: double-sorted oracle quintile portfolios

	Equally Weighted					Value Weighted				
	Comp. Ret.[%]	Ret. [bps/-day]	t-stat	Std. Dev.[%]	Sharpe	Comp. Ret.[%]	Ret. [bps/-day]	t-stat	Std. Dev.[%]	Sharpe
Night Q1	72.59	21.68	23.65	0.70	4.94	16.43	6.04	5.45	0.67	1.14
Night Q2	27.18	9.55	10.92	0.67	2.28	9.44	3.58	3.77	0.67	0.79
Night Q3	15.61	5.76	6.74	0.65	1.41	6.06	2.33	2.62	0.67	0.55
Night Q4	6.99	2.68	3.22	0.63	0.67	4.05	1.58	1.78	0.72	0.37
Night Q5	−7.18	−2.96	−4.00	0.56	−0.84	2.66	1.04	1.19	0.84	0.25
Day Q1	−7.86	−3.25	−2.28	1.08	−0.48	−0.98	−0.39	−0.26	1.16	−0.05
Day Q2	1.41	0.55	0.34	1.13	0.08	1.37	0.54	0.39	1.04	0.08
Day Q3	−1.10	−0.44	−0.23	1.11	−0.06	1.43	0.56	0.42	1.01	0.09
Day Q4	−0.58	−0.23	−0.16	1.09	−0.03	1.88	0.74	0.56	1.01	0.12
Day Q5	1.94	0.76	0.61	0.94	0.13	1.35	0.53	0.39	1.05	0.08

for both equally-weighted and value-weighted portfolios.

The intraday portfolio does not show the same monotonic trend as in the overnight portfolios. There is no evident correlation between the magnitude of volume shocks in a

portfolio and its subsequent intraday performance. Specifically, the mean returns of all equally-weighted portfolios are statistically indistinguishable from zero with the exception of the mean return of the final quintile in the value-weighted portfolios, which is statistically significant. These results provide empirical evidence that volume shocks affect future stock returns when the market is closed, but not when the market is open.

Panel B of Table 5.1 reports that portfolios with stocks that have larger volume shocks yield higher returns and higher Sharpe ratios when the market is closed, but this relationship does not hold when the market is open. For portfolios traded overnight, the variability of the returns of double-sorted equally-weighted portfolios is greater than that of the single-sorted equally-weighted portfolios. The returns of the top two quintiles in double-sorted portfolios are higher than those of the top two quintiles of single-sorted portfolios, while the returns of the bottom three quintiles are lower. In contrast, the returns of the value-weighted portfolios are more uniform across quintiles when they are double sorted than when they are single sorted. Furthermore, the returns and Sharpe ratios of the the first quintile of the value-weighted, double-sorted portfolios are lower than those of the single-sorted portfolios. However, the returns and Sharpe ratios of the remaining four quintiles in the double-sorted portfolios are higher than those of the single-sorted portfolios.

When the market is open, the mean returns of only one of the double-sorted portfolios is statistically different from zero. Overall, the t-statistics of double-sorted intraday portfolios are much smaller than those of the single-sorted portfolios. The t-statistic of only one out of ten intraday double-sorted portfolios is larger than one as opposed to four out of ten for single-sorted portfolios. This observation provides evidence that the relationship between volume shock and intraday return is insignificant, especially after accounting for firm size.

On the other hand, the results for portfolios traded during the market hours on the next day show no significant correlation between volume shocks and stock returns, see Table 5.1. The mean returns of the stocks in the second quintile (by volume shocks) are the least negative at about 1% higher than the market annualized return. The Sharpe ratio of all quintile portfolios differ by less than 0.1. Therefore, the results in Table 5.1 indicate that the relationship between volume shocks and stock returns is more pronounced during the overnight period than those driving the intraday trading session.

Next, we study if the common risk factors explain returns of long-short portfolios sorted by volume shock. Recall that for long-short portfolios we take a long position in the first quintile portfolio and a short position in the fifth quintile portfolio. Table 5.2 presents results from factor regressions of single- and double-sorted-long-short volume shocks portfolios on returns of the market, size, long-term momentum, and short-term reversal portfolios.

Panel A presents results of regressions on returns of single-sorted portfolios. The intercepts of the equally-weighted and value-weighted portfolios are 22.66 basis points and 3.72 basis points per day with t-statistics of 62.48 and 8.31, respectively. Thus, the excess returns of the portfolios sorted on volume shock are statistically significant when the market is closed. On the other hand, the intercepts of the portfolios on intraday returns are much smaller at 1.14 and 2.39 basis points per day for equally-weighted and value-weighted portfolios, with t-statistics of 1.30 and 2.07, respectively. In particular, while the intercept of value-weighted intraday portfolio is one basis point smaller than that of the overnight portfolio, the t-statistic of the intraday portfolio is less than a quarter of that of the overnight portfolio.

The market betas of the equally-weighted and value-weighted overnight portfolios are

Table 5.2: Factor Regressions

Regressions of volume-shock long-short portfolios on returns from the market, size, long-term momentum, and short-term reversal portfolios. Returns are measured daily. Panel A reports results from single-sorted portfolios and Panel B reports results from double-sorted portfolios. "Night Equal" and "Night Value" represent results for equally-weighted and value-weighted portfolios on overnight returns, respectively; and "Day Equal" and "Day Value" represent results for equally-weighted and value-weighted portfolios on intraday returns, respectively. The t-statistics are reported in parentheses. Statistical significance at the 1%, 5%, and 10% level is indicated by ‡, †, and *, respectively. Data are from CRSP.

	Intercept [bps]	β_{mkt}	β_{mom}	β_{rev}	β_{size}	Adj. R^2 [%]
Panel A: Single-sorted portfolios						
Night Equal	$22.66^{\ddagger}$	$0.2217^{\ddagger}$	0.0548^{*}	0.0286	-0.0137	21.0
	(62.48)	(38.86)	(1.87)	(0.69)	(-0.65)	
Night Value	$3.72^{\ddagger}$	$0.1912^{\ddagger}$	$0.1003^{\ddagger}$	$0.2012^{\ddagger}$	$-0.3042^{\ddagger}$	14.6
	(8.31)	(27.27)	(2.78)	(3.92)	(-11.72)	
Day Equal	1.14	$0.9438^{\ddagger}$	$0.0374^{\ddagger}$	-0.0912	-0.0587^{*}	70.1
	(1.30)	(115.92)	(0.85)	(-1.14)	(-1.76)	
Day Value	$2.39^{\dagger}$	$0.8431^{\ddagger}$	0.0653	0.1397	$-0.1960^{\ddagger}$	50.9
	(2.07)	(76.71)	(1.11)	(1.62)	(-4.34)	
Panel B: Double-sorted portfolios						
Night Equal	$23.09^{\ddagger}$	$0.2007^{\ddagger}$	$0.0773^{\ddagger}$	0.0171	-0.0026	18.5
	(64.98)	(35.90)	(2.69)	(0.42)	(-0.13)	
Night Value	$3.43^{\ddagger}$	$0.2038^{\ddagger}$	$0.1383^{\ddagger}$	$0.2582^{\ddagger}$	$-0.3272^{\ddagger}$	14.7
	(7.23)	(27.30)	(3.61)	(4.74)	(-11.87)	
Day Equal	$-4.12^{\ddagger}$	$0.1257^{\ddagger}$	$0.7095^{\ddagger}$	0.0067	$0.0802^{\ddagger}$	12.3
	(-8.67)	(28.39)	(2.98)	(0.19)	(4.41)	
Day Value	-0.75	$0.0431^{\ddagger}$	0.0419	$0.1553^{\ddagger}$	0.0209	0.8
	(-0.96)	(5.96)	(1.08)	(2.73)	(0.70)	

around 0.22 and 0.19, respectively; on the other hand, the market betas for intraday portfolios are around 0.94 and 0.84. For each regression, the t-statistics for the market betas are large. The volume-shock-sorted portfolios are much less correlated to the market when the market is closed than when the market is open. The returns of the overnight portfolios are positively affected by momentum and reversal, but negatively affected by the size factor portfolio returns. These effects are statistically significant for value-weighted portfolios, but not for equally-weighted portfolios. The negative exposure to size is also significant for intraday portfolios. This shows that the market capitalization of stocks and volume shocks of stocks are positively correlated.

Overall, the returns of the long-short portfolios sorted by volume shock are better explained by the four risk factors we include when the market is open than when the market is closed. The adjusted R-squares are 21% and 14.6% for equally-weighted and value-weighted portfolios when the market is closed, and the adjusted R-squares are 70.1% and 50.9% when the market is open. This suggests that there are unobserved risk factors driving the returns of the portfolios sorted by volume shock when the market is closed.

We observe similar results for regressions on returns of double-sorted portfolios in Panel B of Table 5.2. The intercepts of equally-weighted and value-weighted portfolios are around 23 and 3.7 basis points for overnight portfolios, and -4 and -0.75 basis points for intraday portfolios. For equally-weighted and value-weighted portfolios, the t-statistics of the coefficients of portfolio returns are around eight times as large in magnitude when the market is closed than when the market is open. Similar to the results in single-sorted portfolios, the volume-shock-sorted portfolios are positively impacted by the market, momentum, and reversal factors, but are negatively impacted by size when the market is closed. When the market is open, the returns of the double-sorted portfolios are less affected by the market because their market beta decreases from nearly 1 when single-sorted to around 0.1 for double-sorted portfolios; consequently, the adjusted R-square for volume-shock-sorted portfolios decreases significantly for intraday portfolios, but not for overnight portfolios when double sorted on size and volume shock.

Next, Table 5.3 reports results of Fama–MacBeth regressions and pooled regressions. The results for each decade in the 2000s are in Panels A, B, and C.

Panel D of Table 5.3 reports the overall estimates of the risk premia and intercept in the cross-sectional regressions from January 2000 to December 2022. The estimated risk

Table 5.3: Fama–MacBeth and Pooled Regressions

Fama–MacBeth regressions of intraday and overnight returns on betas from volume shocks and market returns. Returns are measured during the Day, from open-to-close, and during the Night, from close-to-open. Betas are estimated daily on 252-day rolling window time series regressions for each stock. For pooled regressions, standard errors are clustered by date. Panels A, B, C, D report results from 2000 to 2009, 2010 to 2019, 2020 onwards, and overall results from 2000 to 2022, inclusive, respectively. The t-statistics are reported in parentheses. Statistical significance at the 1%, 5%, and 10% level is indicated by ‡, †, and *, respectively. Data are from CRSP.

	Fama–MacBeth regressions				Pooled regressions				
	Avg. ξ_0	Avg. ξ_1	Avg. ξ_2	Avg. R^2	β_{VS}	β_{Mkt}	Day	Day·β_{VS}	R^2
Panel A: 2000-2009									
Night	$-0.0007^{\ddagger}$	$0.3265^{\ddagger}$	$0.0010^{\ddagger}$	1.58	$0.3869^{\ddagger}$	$0.0007^{\ddagger}$	-0.0002^{*}	$-0.4928^{\ddagger}$	0.04
	(−7.47)	(23.06)	(5.39)		(16.19)	(−1.73)	(2.89)	(−14.32)	
Day	$0.0015^{\ddagger}$	$-0.0846^{\ddagger}$	$-0.0011^{\ddagger}$	1.43					
	(10.93)	(−5.60)	(−5.55)						
Panel B: 2010-2019									
Night	$0.0003^{\ddagger}$	$0.4213^{\ddagger}$	$0.0003^{\dagger}$	1.81	$0.3828^{\ddagger}$	−0.0001	$-0.0007^{\ddagger}$	$-0.5701^{\ddagger}$	0.07
	(7.37)	(33.83)	(2.14)		(27.20)	(−3.30)	(−0.70)	(−26.38)	
Day	$0.0004^{\ddagger}$	$-0.2137^{\ddagger}$	$-0.0005^{\ddagger}$	1.33					
	(4.43)	(−16.68)	(−3.10)						
Panel C: 2020-2022									
Night	$0.0019^{\ddagger}$	$0.4198^{\ddagger}$	-0.0008^{*}	3.07	$0.4728^{\ddagger}$	-0.0005^{*}	$-0.0019^{\ddagger}$	$-0.5201^{\ddagger}$	0.17
	(12.15)	(13.13)	(−1.82)		(12.26)	(−1.91)	(−3.33)	(−9.07)	
Day	$-0.0009^{\ddagger}$	$-0.0969^{\ddagger}$	0.0001	2.62					
	(−2.43)	(−3.74)	(0.31)						
Panel D: 2000-2022									
Night	$0.0001^{\ddagger}$	$0.3823^{\ddagger}$	$0.0004^{\ddagger}$	1.89	$0.3866^{\ddagger}$	$-0.0002^{\ddagger}$	−0.0002	$-0.4867^{\ddagger}$	0.04
	(2.77)	(41.48)	(3.79)		(23.89)	(−2.72)	(−1.27)	(−20.55)	
Day	$0.0007^{\ddagger}$	$-0.1451^{\ddagger}$	$-0.0006^{\ddagger}$	1.55					
	(8.02)	(−15.71)	(−5.31)						

premia for volume shocks are 0.3823 per day with t-statistics of 41.48 when the market is closed; the volume shock risk premia is −0.1451 per day with t-statistic −15.71 when the market is open. These results show that the risk premia for volume shock are strongly and significantly positive when the market is closed and significantly negative when the market is open. The estimated risk premia of volume shock are large in magnitude because the value of the volume shock beta is much smaller than the beta of the market;[7] while

[7]Volume shock of individual stocks can be large in value. It is common for the volume shock of a stock

the average of market beta is around 0.85, volume shock beta are only around 15 basis points, on average. Consequently, the volume shock term in equation (5.6) contributes to around 5 basis points to the cross-sectional return of stocks, on average. An increase of 10 basis point in volume shock beta increases the return of the stock by 3.8 basis points when the market is closed and it decreases it by around 1.5 basis points when the market is open. The risk premia for the market estimated in the Fama–MacBeth regression are also positive overnight and negative intraday, consistent with findings in Hendershott et al. (2020).

Findings from Fama–MacBeth regressions in Panel D of Table 5.3 are confirmed by pooled regressions. We use a singe pooled panel regression as specified in (5.7) to compute the difference in volume shock coefficients between periods when the market is closed and when the market is open. Standard errors are clustered at the day level for the pooled regressions. The regression coefficient on $Day \cdot \beta_{VS}$ captures the difference in the volume shock coefficients from overnight to intraday. The volume shock beta coefficient is 0.3866 with t-statistic 23.89, and the estimate of the coefficient of the term $Day \cdot \beta_{VS}$ is -0.4867 with t-statistic -20.55. These findings are consistent with the findings of the Fama–MacBeth regressions. When the market is closed ($D = 0$), the volume shock beta coefficients are almost identical to those we find in the Fama–MacBeth regression; and when the market is open ($D = 1$), the difference between the volume shock beta coefficient and the $Day \cdot \beta_{VS}$ is $0.3866 - 0.4867 = -0.10$, which is very close to the estimated risk premia for volume shock when the market is open. The coefficients for the "Day" dummy are statistically insignificant, and the R-square of the regression is 0.04%.

Panels A, B, C of Table 5.3 report the evolution of risk premia and intercepts in the

to be between 10 and 15.

Fama–MacBeth cross-sectional regressions and pooled regressions during different periods in the history. Approximately, results in Panel A cover the time before and during the financial crisis; results in Panel B cover the time after the financial crisis and before Covid; and results in Panel C cover the time during and after Covid.

Similar to the overall results in Panel D, in each of the subsamples, the estimated risk premia for volume shock are positive and statistically significant when the market is closed, and they are negative and statistically significant when the market is open. Notably, when the market is closed, the estimated risk premia have smaller magnitude as well as t-statistics than when the market is open. The risk premia for volume shock are the largest in magnitude from 2010 to 2019 both intraday and overnight. Its magnitude is smallest overnight from 2000 to 2009 and intraday from 2020 onwards. Interestingly, the estimated market risk premia is insignificant in the third subsample during Covid and the intercept also becomes larger in magnitude and in t-statistics compared to before Covid, which may suggest that Covid changed the market dynamics fundamentally.

The pooled regressions on each of the subsamples also confirm the results in the subsample Fama–MacBeth regressions. The coefficients on volume shock beta are large in magnitude and are significantly positive, and the coefficients on the interaction term $Day \cdot \beta_{VS}$ are large in magnitude and are significantly negative. The magnitude of the volume-shock beta coefficient is similar to that of the overnight risk premia term in the Fama–MacBeth regressions, while the magnitude of the difference between the volume shock term and interaction term is similar to that of the intraday Fama–MacBeth risk premia. These results show that the positive correlation between volume shock and stock return overnight and the negative correlation between volume shock and stock return intraday are statistically significant and persistent regardless of market conditions.

Table 5.4: Market Regression

Regressions of market returns on returns from the volume shock, reversal, momentum, and size sorted portfolios. Returns of the portfolios and the market are measured daily during the Day, from open-to-close, and during the Night, from close to open on value-weighted portfolios. Panel A reports results from portfolios formed during the day; panel B reports results from overnight portfolios. The t-statistics are reported in parentheses. Statistical significance at the 1%, 5%, and 10% level is indicated by ‡, †, and *, respectively. Data are from CRSP.

	Intercept	β_{vs}	$\beta_{\text{mkt,t-1}}$	β_{mom}	β_{rev}	β_{size}	Adj. R^2 [%]
Panel A: Night Portfolios							
With VS	-0.00003	$0.6843^{\ddagger}$	$-0.0991^{\ddagger}$	-0.0047	-0.0024	$-0.3499^{\ddagger}$	17.5
	(-0.42)	(30.83)	(-8.27)	(-0.07)	(-0.03)	(-7.02)	
No VS	$0.0003^{\ddagger}$	\	$-0.1207^{\ddagger}$	0.0985	$0.1791^{\dagger}$	$-0.6292^{\ddagger}$	3.9
	(3.36)		(-9.35)	(1.33)	(1.71)	(-11.89)	
Panel B: Day Portfolios							
With VS	0.0002	$0.2699^{\ddagger}$	$-0.0531^{\ddagger}$	$-0.1539^{\dagger}$	$0.3056^{\ddagger}$	$-0.3194^{\ddagger}$	3.9
	(1.29)	(11.88)	(-4.08)	(-2.31)	(3.13)	(-6.27)	
No VS	0.0002	\	$-0.0719^{\ddagger}$	$-0.1480^{\dagger}$	$0.3574^{\ddagger}$	$-0.3167^{\ddagger}$	1.5
	(1.16)		(-5.49)	(-2.19)	(3.62)	(-6.14)	

Next, we study if adding the returns of the value-weighted long-short volume shock portfolio in (5.8) significantly improves the fit of the regression, i.e., we explore if volume shock explains the market level intraday and overnight returns.

Panels A and B of Table 5.4 report the regressions of the returns of the market on various risk factors. Panel A reports the regressions on overnight market returns; Panel B reports the regressions on intraday market returns. In each panel, we use "No VS" to denote the regression without the long-short volume shocks portfolio and "With VS" to denote the regression that includes the long-short volume shocks portfolio return as an independent variable.

For the regressions on overnight returns in Panel A, including the volume-shock long-short portfolio in the regression significantly increases the adjusted R-square of the model from

3.9% to 17.5%. The coefficient on the volume-shocks portfolio is 0.6843 with t-statistics 30.83; meanwhile, including the volume shocks portfolio in the regression on intraday market portfolios in Panel B only increases the adjusted R-square of the regression from 1.5% to 3.9% and the magnitude of the coefficients for volume shocks is 0.2699 with t-statistics 11.88. These observations show that the impact of volume shock portfolio on market return is positive, more pronounced, and statistically significant when the market is closed but smaller in magnitude and t-statistics when the market is open.

The market portfolio also exhibits statistically significant reversal patterns. When the market is closed, the market portfolio exhibits approximately 10% reversal from the previous overnight trading session, and when the market is open, the reversal pattern is approximately 7%.

The only other statistically significant factor that explains the overnight return of the market portfolio is size. After weighting by value, the market portfolio is negatively affected by the returns of the small minus large size portfolio when the market is closed. Similar to the observations for volume shock, when the market is closed, the effect of size on market portfolio is larger in magnitude and in absolute value of t-statistics than when the market is open. This observation implies that similar to volume shock, the behaviour of many other risk factors may be different when the market is closed than when the market is open, consistent with the observations in Lou et al. (2019).

Overall, including the term accounting for volume shocks helps to explain the returns of the market when the market is closed more than when the market is open. The effect of volume shocks on market returns is statistically significant, positive, and large in magnitude when the market is closed; i.e., the larger the return of the volume-shock long-short portfolio,

the larger the overnight return of the market. This evidence, together with the previous results in Table 5.3, shows that volume shocks partly account for the variabilities in cross-sectional stock returns and market returns, and this effect is particularly significant when the market is closed.

5.4.2 Prediction of Volume shocks and trading strategies

Recall that the portfolios we use to verify that volume shock positively affects future overnight returns are not tradable because volume shock is only known after the close auction. To circumvent this issue, we use machine learning models to predict volume shocks for each stock and trade long-only portfolios at the close auction to understand how much economic significance one can obtain from being able to predict volume shock better.

Recall that we construct long-only portfolios instead of long-short portfolios because stocks with low volume shocks may be too iliquid to trade. Therefore, one baseline is the first volume-shock quintile in Table 5.5; recall that it is a value-weighted long only portfolio. Recall, we refer to it as the "oracle" portfolio because it trades on volume shock that is not yet known at the start of the close auction when one makes the trade. We measure the performance of the oracle portfolio in the out-of-sample period from the start of 2016 to the end of 2022 as a baseline for our machine learning based portfolios. After we sort stocks on predicted values of volume shock, we construct long-only double-sorted value-weighted portfolios in the same way as those discussed in Table 5.5. The performance of the oracle portfolio serves as the theoretical maximum of any long-only portfolios that are sorted with predictions of volume shock. We explore if better predictions of volume shock result in portfolios whose performance are closer to the oracle portfolio.

Table 5.5 reports the in-sample and out-of-sample R-square value, mean-squared-error (MSE), and the performance of the machine learning based long-only portfolios. We use annualized returns and Sharpe ratio of the value-weighted portfolio to evaluate the performance of the portfolios we construct.

Results in Table 5.5 show that machine learning methods perform better in predicting volume shock both in-sample and out-of-sample. The in-sample MSE of LightGBM and TabNet are 74% and 87% of the MSE of the Ordinary Least Squares (OLS) model, respectively. This gap narrows out-of-sample but remains significant; the MSE of LightGBM and TabNet are 88% and 81% of the MSE of OLS out-of-sample. This gap is more pronounced when looking at R-square values; the R-square of machine learning algorithms is almost twice the value of those in the linear regression methods. On the other hand, after grid searching for best hyper parameters, Lasso underperforms OLS both in and out-of-sample by around 6% in terms of out of sample MSE and 2% in terms of out-of-sample R-square.

These observations show that machine learning algorithms better capture complex patterns in the dynamics of stock volume shocks and make more accurate predictions than linear methods.

Table 5.5: Performance of various machine learning models for predicting volume shocks on individual stock levels. In-sample training data runs from 2000 and January 2016, and out-of-sample test data runs from 2016 onwards. MSE of the models are normalized so that the MSE for the linear model is 1 in sample. The daily return and Sharpe ratios of the model-induced portfolios are computed on out-of-sample periods.

Model	In-Sample R-square	In-Sample MSE	Out-of-Sample R-square	Out-of-Sample MSE	Ann. Return	SR
Linear	0.142	1.000	0.178	0.961	16.03	1.00
Lasso	0.132	1.029	0.159	1.018	14.25	0.90
LightGBM	0.364	0.741	0.277	0.846	16.49	1.05
TabNet	0.266	0.871	0.333	0.780	18.03	1.08
Oracle	\	\	\	\	17.76	1.11

Next, we explore if better predictions of volume shock lead to more profitable portfolios. Table 5.5 reports the return and Sharpe ratios of portfolios built with volume shock prediction models compared with the baseline "oracle" portfolio. Indeed, returns and Sharpe ratio of the portfolios built with machine learning models tend to be higher than portfolios built with linear models. Specifically, the OLS model reports a Sharpe ratio of 1 and annualized return of 16%, corresponding to 91% and 90.3% of the oracle portfolio, respectively. In contrast, the top-performing machine learning model achieves a Sharpe ratio of 1.08 and an annualized return of 18%, which are 97.3% and 101.5% of the oracle portfolio.

These findings suggest that one can build a portfolio that achieves around 90% of the performance of the oracle portfolio with the simplest models, and with better models to predict volume shock, one can achieve around 97% of the performance of the oracle portfolio. This finding assures that the economic significance of volume shocks are accessible to investors.

While we observe a monotonic relationship between R-square and portfolio returns and Sharpe ratios, this relationship is not linear. Compared to the Lasso regression portfolios, the R-square and Sharpe ratios of OLS are 12% and 11% higher, respectively, but when comparing OLS and the machine learning models, one needs a much larger improvement in R-square to achieve similar levels of improvement in Sharpe ratios. For example, the out-of-sample R-square of TabNet is 87% higher than that of OLS, but the Sharpe ratio of the portfolio created from TabNet is only 8% larger than that of the OLS portfolio. This finding provides evidence that the relationship between predictive errors and portfolio performance is convex; when the R-square of the model is low, a small improvement in predictive power can lead to a large increase in Sharpe ratio, but when the R-square of

the model is high, a better predictive model contributes marginally to the performance of portfolios.

6 | Correlation Matrix Clustering for Statistical Arbitrage Portfolios

6.1 Introduction

Statistical arbitrage encompasses investment strategies that use statistical and quantitative methods to identify and exploit temporal price deviations among a group of similar assets. An example of a classical statistical arbitrage strategy is pairs trading Elliott et al. (2005); Cartea and Jaimungal (2016); Cartea et al. (2019), which takes a long position in one security and a short position in another security with the expectation that the spread between their prices will revert back to an anticipated level. While works such as Bergault et al. (2022); Bertram (2010) use the Ornstein-Uhlenbeck process to model stock prices, we focus on a model agnostic statistical arbitrage framework that consists of two steps, (1) identify a group of assets that share similarities, and (2) construct an arbitrage portfolio within the group of assets.

In this paper, we propose a framework where we employ graph clustering algorithms to identify groups of correlated stocks that co-move. Then, we construct mean-reverting arbitrage portfolios within each cluster to evaluate if the clustering methods enable statistical arbitrage strategies that deliver economically significant profits.

In the first step, we compute market residual returns, which are given by the difference between the stock returns and the product of the stock's CAPM Black et al. (1972) equity beta and the market return.[1] Next, we compute the correlation matrix of residual returns,

[1] In this paper, we use the return of the *SPY* ETF as a proxy for market returns.

interpret it as a weighted signed network, and use graph clustering algorithms to partition the stocks into groups such that on average, the correlation between stocks that are in different groups is low and the correlation between stocks in the same group is high. In this paper, we employ five clustering algorithms to construct statistical arbitrage portfolios. The clustering algorithms include two variants of SPONGE clustering Cucuringu et al. (2019a), a modified variant of Spectral clustering Ng et al. (2002), and two variants of the Signed Laplacian clustering Kunegis et al. (2010).

In the second step, we employ a rolling window to identify stocks whose returns are above and are below the mean returns of the cluster, which we label "previous winners" and "previous losers", respectively. Next, we construct a contrarian portfolio that consists of a long position on the previous losers and a short position on the previous winners within each cluster. We use this portfolio to evaluate if the stocks in each cluster exhibit mean-reversion patterns, i.e., the returns of stocks in each cluster revert to the mean return of the cluster.

There is an active strand of literature that applies clustering methods in portfolio management. Also, there are two main approaches to construct portfolios after grouping securities into clusters, which we detail below.

One approach uses all stocks in each identified cluster to construct mean-variance Markowitz portfolios Markowitz (1952). For example, León et al. (2017); Tola et al. (2008) first cluster the correlation matrix, then group the stocks according to the corresponding entries in the correlation matrix, and afterwards construct Markowitz portfolios within each cluster. Alternatively, Gatta et al. (2023) uses regression coefficients of asset returns on various factors to cluster securities, and then constructs a variance minimizing portfolio in each

cluster.

A second approach in the literature clusters the correlation matrix as in the first approach, and then selects one asset from each of the clusters to construct a single Markowitz portfolio. For example, Tolun Tayalı (2020) constructs a Markowitz portfolio with the medoids of each cluster. Other lines of work that follow this second approach employ various clustering methods and selection mechanisms to identify the representative stock in each cluster. For example, Wang et al. (2022) selects the stock with the lowest volatility within each cluster, and Tang et al. (2021) selects the stock with the highest Sharpe ratio in each cluster.

To the best of our knowledge, ours is the first work that applies clustering algorithms in the design of statistical arbitrage strategies.

Our approach draws from the literature on clustering algorithms. In particular, we use signed clustering algorithms that handle negative weights because the correlation matrix of residual returns we employ consists of both positive and negative entries, so one cannot apply many of the classical graph clustering algorithms proposed in the literature. Previous lines of work, including Aghabozorgi et al. (2015); Focardi (2005); Ziegler et al. (2010); Pavlidis et al. (2006), use signed clustering methods to analyze financial time series such as macroeconomic variables and time series of large baskets of stock returns.

6.2 Mathematical Model and Problem Setting

6.2.1 Signed & Directed Graph Clustering

Clustering is a widely used technique in data analysis. A clustering algorithm identifies groups of nodes in a network that exhibit similar behavior or features. In this paper, we focus on a strand of clustering algorithms that operates on the spectrum of suitably defined matrix operators that are built directly from the data. Such methods, often referred to as "spectral methods", are the subject of a growing body of literature in the last decade, mainly motivated by their computational efficacy, robustness to noise, and amenability to theoretical guarantees that rely on results from the random matrix theory and matrix perturbation literature.

This section introduces the clustering methods we use to construct statistical arbitrage portfolios. In particular, it introduces clustering methods that operate on signed networks (i.e., with adjacency matrices that are symmetric and contain both positive and negative entries).

We partition a signed network into K clusters such that most edges within clusters are positive, and most edges across clusters are negative. To achieve this, we seek a partition that minimizes the number of violations; i.e., negative edges within each cluster and positive edges across clusters, as illustrated in Figure 6.1.

6.2.1.1 Spectral Clustering

Spectral clustering Ng et al. (2002) is one of the simplest and most popular spectral methods that clusters a network based on the adjacency matrix of the network. Here, we

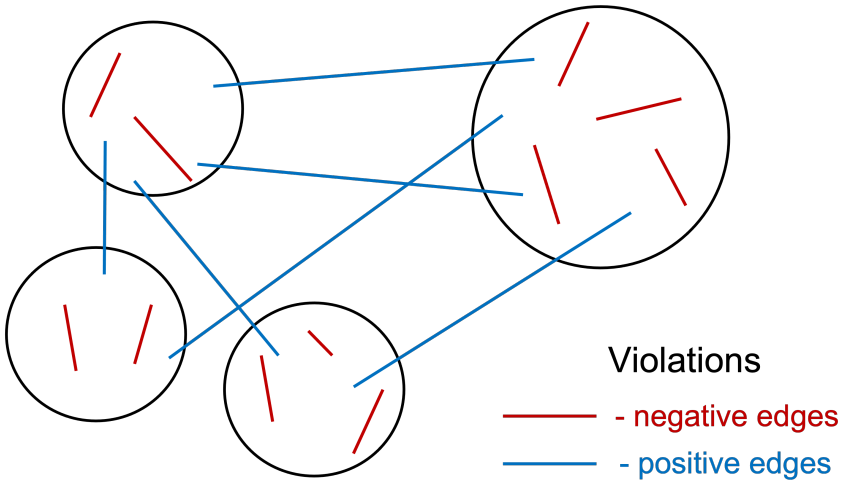


Figure 6.1: Signed clustering minimizes the number of violations in the constructed partition. A violation, as in this figure, is when there are negative edges in a cluster and positive edges across clusters.

apply spectral clustering on a correlation matrix of stock market residual returns.

In our approach, the input data is the correlation matrix of stock market residual returns. Each stock corresponds to a node in the network, and the correlation between stock returns represents the co-movement similarity between the nodes. We use the correlation matrix to build the graph Laplacian matrix $\mathbf{L}$, defined as the difference between the degree matrix $\mathbf{D}$ and the adjacency matrix $\mathbf{A}$ of the similarity graph. The degree matrix $\mathbf{D}$ is a diagonal matrix that captures the degree or total strength of connections for each node in the graph, while the adjacency matrix $\mathbf{A}$ encodes the pairwise similarities between nodes, as determined by the edge weights. The Laplacian matrix measures the difference between the sum of similarities connecting a node to the rest of the network, and it measures the node's total strength of connections. For the Spectral clustering algorithm, the entries of the similarity matrix must be positive; therefore, we take the absolute value of the correlation matrix and use this modified correlation matrix as the adjacency matrix in the Spectral clustering algorithm. In Knyazev (2017), the author uses the standard graph Laplacian matrix to perform Spectral clustering ignoring that some of the edge weights are negative;

however, later works, including Cucuringu et al. (2019a), report poor performances of this approach. Therefore, in this paper, we employ the unsigned Spectral clustering algorithm which considers the absolute value of the correlation matrix.

Formally, the Laplacian matrix $\mathbf{L}$ is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}\,, \tag{6.1}$$

and recall that $\mathbf{D}$ is the diagonal degree matrix and $\mathbf{A}$ is the adjacency matrix. The diagonal elements of $\mathbf{D}$ are the sums of the weights (similarities) of the edges that are connected to each node, while the off-diagonal elements of $\mathbf{A}$ represent the pairwise similarities between nodes. In our case, $\mathbf{A}$ is the absolute value correlation matrix and $\mathbf{D}_{ii} = \sum_{i=1}^{n} \mathbf{A}_{ij}$.

Next, we find the K smallest eigenvectors of the Laplacian matrix to obtain a low-dimensional embedding, from which we subsequently extract K clusters. These K eigenvectors, which correspond to a K-dimensional Euclidean space, are the input for k-means++ clustering that partitions the nodes into disjoint clusters.

6.2.1.2 Signed Laplacian Clustering

The Signed Laplacian clustering algorithm operates on the Signed Laplacian matrix, in contrast to the unsigned Laplacian in the case of Spectral clustering. Mathematically, the graph Signed Laplacian is defined in a similar way to the graph Laplacian. The differences are that the adjacency matrix $\mathbf{A}$ can take negative values and that the degree matrix $\overline{\mathbf{D}}$ is $\overline{\mathbf{D}}_{ii} = \sum_{i=1}^{n} |\mathbf{A}_{ij}|$.

In Kunegis et al. (2010), the authors use the spectrum of the Signed graph Laplacians to perform clustering. Specifically, they extend the ratio cut and normalized cut functions

from the unsigned literature to signed graphs and use the Signed Laplacian matrix to perform clustering.

To ensure the Signed Laplacian matrix is symmetric and positive semi-definite, the algorithm normalizes the Signed Laplacian in two alternative ways. One, the random-walk normalized Laplacian is constructed as $\overline{\mathbf{L}}_{\text{rw}} = \mathbf{I} - \overline{\mathbf{D}}^{-1}\,\mathbf{A}$. Two, the symmetric normalized graph Laplacian is defined as $\overline{\mathbf{L}}_{\text{sym}} = \mathbf{I} - \overline{\mathbf{D}}^{-1/2}\,\mathbf{A}\,\overline{\mathbf{D}}^{-1/2}$. We employ both normalization methods in our empirical study.

After normalizing the Signed Laplacian, the algorithm solves an optimization problem on the normalized cut functions. First, it computes the K smallest eigenvectors of the normalized Signed Laplacian and then performs k-means++ clustering on the Euclidean space of eigenvectors.

6.2.1.3 SPONGE — a generalized eigenproblem

SPONGE (Signed Positive Over Negative Generalized Eigenproblem) is a generalized eigenvalue formulation of the signed clustering problem, which outperforms many benchmarks in the literature Cucuringu et al. (2019a). The algorithm is particularly effective when the number of clusters K is large or when the underlying graph is very sparse.

The algorithm first decomposes the adjacency matrix $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, with $\mathbf{A}^+_{ij} > 0$ and $\mathbf{A}^-_{ij} > 0$. Next, it constructs two Laplacian matrices $\mathbf{L}^+$ and $\mathbf{L}^-$, with corresponding diagonal degree matrices $\mathbf{D}^+$ and $\mathbf{D}^-$ based on $\mathbf{A}^+$ and $\mathbf{A}^-$, respectively. The algorithm minimizes the ratio between the positive cuts and negative cuts, while adding regularization terms to promote clusterizations that avoid small-sized clusters. The approach extends to multiple clusters, and, under appropriate assumptions and changes of variables, it leads to a generalized eigenvalue problem which can be solved efficiently using

pre-conditioners Knyazev (2001).

In practice, the SPONGE algorithm finds the K smallest generalized eigenvectors of $(\mathbf{L}^+ + \tau^- \mathbf{D}^-, \mathbf{L}^- + \tau^+ \mathbf{D}^+)$, where τ^+, $\tau^- > 0$ are regularization parameters. The algorithm then performs k-means++ clustering on the induced K-dimensional Euclidean space.

We also employ the variant SPONGE$_{sym}$ of the SPONGE algorithm, which relies on the symmetric graph Laplacian $\overline{\mathbf{L}}_{\text{sym}}$. This variant first finds the smallest K generalized eigenvectors of $(\mathbf{L}^+{}_{sym} + \tau^- \mathbf{I}, \mathbf{L}^-{}_{sym} + \tau^+ \mathbf{I})$, where $\mathbf{L}^+{}_{sym} = (\mathbf{D}^+)^{-1/2} \mathbf{L}^+ (\mathbf{D}^+)^{-1/2}$ is the symmetric Laplacian of $\mathbf{A}^+$ (and similarly for $\mathbf{L}^-{}_{sym}$). This symmetric Laplacian is useful for networks with skewed degree distributions. Under suitably defined signed stochastic block models, Cucuringu et al. (2019a, 2021) provide theoretical cluster recovery guarantees (upper bound on the misclustering rate) for the SPONGE family of algorithms as a function of the noise and edge sparsity levels.

6.2.2 Portfolio Construction

We construct K zero-cost statistical arbitrage portfolios within a universe of N stocks, where K is the number of clusters we identify. There are four steps:

1. Data pre-processing,
2. Group stocks into disjoint clusters,
3. Identify a collection of stocks within each cluster such that a linear combination of them is likely to mean-revert to zero,
4. Assign portfolio weights to the selected stocks in each cluster.

6.2.2.1 Data Pre-processing

First, we compute the market residual return $R_{i,t}^{res}$ of stock i at time t, which is given by

$$R_{i,t}^{res} = R_{i,t} - \beta_i \, R_{mkt,t} \, , \tag{6.2}$$

and where $R_{i,t}$ is the raw return of stock i at time t, β_i is the beta coefficient of stock i, which measures its sensitivity to market movements, and $R_{mkt,t}$ is the market return at time t. In our empirical study below, the market is the SPY ETF and we use a 60 trading day rolling window to estimate β and compute the market residual return.

The market residual return represents the component of the stock's return that is not explained by overall market movements, i.e., the idiosyncratic dynamics of each stock. In our case, we use the market residual return to focus on the portion of the stock returns that are specific to the stock themselves and to study the commonalities in their idiosyncratic dynamics.

After computing the market residual return of each stock, we construct the correlation matrix of market residual returns, which we use as the input of the later steps of portfolio construction.

Suppose at time T we want to construct the correlation matrix of market residual returns for N stocks, we first obtain the market residual return of stocks from time $T - W$ to $T - 1$, inclusive. Next, we organize these residual returns into a matrix $\mathbf{R^{res}}$ of dimension w days by N_t stocks. Each element $R_{t,i}^{res}$ in this matrix corresponds to the residual return of stock i on day t.

Then, we compute the entries of the correlation matrix $\mathbf{C}$ as follows

$$C_{i,j} = \frac{\sum_{t=T-w}^{T-1} \left(R_{t,i}^{res} - \overline{R}_i^{res}\right) \left(R_{t,j}^{res} - \overline{R}_j^{res}\right)}{(w-1)\,\sigma_i\,\sigma_j}\,, \tag{6.3}$$

where $\overline{R}_i^{res}$ denotes the mean of the residual returns of stock i, σ_i and σ_j are the standard deviations of returns for stocks i and j over the w days. The resulting correlation matrix $\mathbf{C}$ is of size $N \times N$ and contains the pairwise correlation coefficients between all stocks in the matrix $\mathbf{R^{res}}$.

6.2.2.2 Group stocks into clusters

We employ the clustering algorithms outlined in Section 6.2.1 to partition the correlation matrix of stock market residual returns into K distinct and non-overlapping clusters. Next, we group stocks according to the clusters we obtain from the correlation matrix which groups stocks based on their residual returns while remaining agnostic to market factor moves.

6.2.2.3 Identify stocks to trade

After computing the clusters, we extract arbitrage signals for each stock. Within each cluster, we compute the mean raw return of stocks over a lookback period of w days, and we measure the cumulative deviation of each stock's raw returns from the cluster mean over the past w days.[2]

Consider the returns of stocks $R_{1,t}, ...R_{j_n,t}$ in cluster j, define the cluster mean return at time T over the lookback period of w days as

$$\overline{R}_{j,t} = \frac{1}{j_n} \sum_{i=1}^{j_n} R_{i,t}\,. \tag{6.4}$$

[2]Empirical results show that our portfolio is market-neutral despite using raw returns.

Recall that we identify stocks that outperform the cluster mean over the past w days as previous winners, and stocks that underperform the cluster mean over the past w days as previous losers. In particular, we set a threshold p such that stocks whose returns cumulatively deviate by more than a threshold p from the cluster mean are believed to be more likely to revert back to the cluster mean. We expect the previous winners to revert down to the cluster mean and the previous losers to revert up to the cluster mean over the next T days.

We identify stock j_i with return $R_{j_i,t}$ in cluster j as a previous winner if:

$$\sum_{t=T-w}^{T-1} (R_{j_i,t} - \overline{R}_{j,t}) > p,$$

and as a previous loser if:

$$\sum_{t=T-w}^{T-1} (R_{j_i,t} - \overline{R}_{j,t}) < -p.$$

6.2.2.4 Assign weights to stocks

After identifying the previous winners and the previous losers in each cluster, we assign weights to these stocks and execute a contrarian trading strategy over the next ℓ days.

Within each cluster, we short-sell the previous winners, while simultaneously initiating long positions on previous losers. Portfolio weights are the same for all stocks. Specifically, we normalize the portfolio weights such that the total dollar value of both long and short positions sums up to one within each cluster, which guarantees a zero-cost arbitrage portfolio at inception. For example, if there are two previous winners and four previous losers in a cluster, we construct a portfolio that short-sells fifty cents on each of the pre-

vious winners and bets twenty-five cents on each of the previous losers. This approach aligns with our objective of capturing mean-reversion opportunities, while maintaining a cost-neutral trading strategy.

At the end of ℓ days, we re-balance the portfolio. We first re-compute the correlation matrix of market residual returns over the w day lookback window; then re-compute the clusters and re-construct the portfolio.

To manage risk and optimize performance, we introduce a stop-win threshold at q. Should our portfolio realize a return of q before the completion of the ℓ days, we interpret this as evidence of successful mean-reversion. In that case, we immediately re-balance the portfolio in the same way as if the ℓ-day trading period had ended. With this stop-win mechanism, the portfolio is exposed to profitable mean-reversion events, while mitigating downside risk.

Below, we discuss the choice of the number of clusters K we find to construct the portfolios.

6.2.3 Choosing the Number of Clusters

Determining the number of clusters to partition the network is not straightforward; the literature explores various approaches. Here, we employ methods from random matrix theory and standard statistical analysis to dynamically determine how many clusters to extract every time we construct the arbitrage portfolios.

Consider a universe of N stocks over T days, and store the market residual returns of these stocks in an T by N matrix denoted $\mathbf{X}$. Let $\mathbf{C} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ be the N-by-N empirical correlation matrix of $\mathbf{X}$. The Marchenko–Pastur theorem characterizes the limiting behavior of the eigenvalues of $\mathbf{C}$. It states that, if the entries of $\mathbf{X}$ are independent identically

distributed random variables with mean 0 and finite variance, then as $N, T \to \infty$, with ratio $\rho = N/T$ fixed, the empirical distribution of the eigenvalues of $\mathbf{C}$ converges to the Marchenko–Pastur distribution.

The Marchenko–Pastur distribution characterizes the limiting distribution of eigenvalues of Wishart matrices, and is defined by the density function

$$f(\lambda) = \begin{cases} \frac{\sqrt{(\lambda_+ - \lambda)\,(\lambda - \lambda_-)}}{2\,\pi\,\lambda\,\sigma^2\,\rho}, & \text{for } \lambda \in [\lambda_-, \lambda_+]\,, \\ 0, & \text{otherwise,} \end{cases} \tag{6.5}$$

where $\lambda_- = (1 - \sqrt{\rho})^2$ and $\lambda_+ = (1 + \sqrt{\rho})^2$.

To determine the number k of eigenvalues that provides the dimension of the low-dimensional embedding, we select the eigenvalues of the correlation matrix that exceed the threshold λ^+, which are the eigenvalues associated with dominant factors or patterns in the stock returns. Therefore, k and the number of clusters K are the same, as detailed in Section 6.2.1.

An alternative, and more classical approach, is to consider the total variance explained by the eigenvalues and select the number of largest eigenvalues needed to account for a specific proportion P of the total variance. Specifically, sort the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$ of the correlation matrix $\mathbf{C}$ in decreasing order, such that λ_1 is the largest eigenvalue and λ_N is the smallest. To determine the number of eigenvalues needed to account for a proportion P of the total variance of $\mathbf{C}$, compute the cumulative sum of the eigenvalues and divide by the sum of all eigenvalues, i.e.,

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{N} \lambda_i} \geq P\,. \tag{6.6}$$

Here, k is the number of eigenvalues required to reach or exceed the threshold P, which is also the number of clusters used to partition the network.

For both methods above, we recompute the desired number of clusters every time we construct the mean-reverting portfolio, and use a twenty-day lookback window on stock returns to determine the number of clusters. As a benchmark, we also include performance results when the number of clusters is fixed to 30.[3]

6.2.4 Benchmarks and Evaluation Criterion

We compare the performance of each cluster-driven portfolio with two benchmarks. The first benchmark is the SPY ETF. The second benchmark is an arbitrage portfolio based on the Fama–French 12 industry classifications, which is constructed by building statistical arbitrage portfolios within each of the Fama–French 12 industries in the same way as our cluster-driven portfolios. The second benchmark compares intra-cluster mean-reversion effects to the mean-reversion effect discovered by the cluster-driven portfolios.

To evaluate the performance of our portfolios, we use metrics including annualized return, Sharpe ratio, and Sortino ratio Sortino Frank A. (1994). The Sharpe ratio measures the risk-adjusted return. It considers the market residual return generated per unit of standard deviation of returns as a measure of risk. A higher Sharpe ratio indicates better risk-adjusted performance, i.e., a higher return relative to the amount of risk taken.[4]

On the other hand, the Sortino ratio focuses on the downside risk of the portfolio with the intuition that high upside standard deviation does not negatively impact portfolios and is not a concern for investors. It takes into account only the standard deviation of negative

[3]The dynamic algorithms pick 10 to 20 sectors in most days; we choose 30 clusters to differentiate from the number that we choose dynamically.

[4]For convenience, in this paper, the risk-free rate is set to zero.

returns, providing a measure of risk-adjusted returns targeting the downside volatility, and is defined as

$$\text{Sortino Ratio} = \frac{\text{Portfolio Return} - \text{Risk-Free Rate}}{\text{Downside Deviation}}. \tag{6.7}$$

Here, the downside deviation represents the standard deviation of negative returns. A higher Sortino ratio implies better risk-adjusted performance because it indicates higher returns relative to the downside volatility of the portfolio.

6.3 Empirical Results

6.3.1 Data

Stock price data are from the Center of Research in Security Prices (CRSP) daily returns database. The sample period is from January 2000 to December 2022. We include stocks listed on the NYSE, Amex, and NASDAQ exchanges. For each trading day, to ensure that the trading positions we take are realistic, we only include stocks in the top 25 percentile of market capitalization, which is defined as the product of the price of stock at the end of the day (i.e., close price) and the number of shares outstanding. The stock universe we include consists of around 600 stocks in each trading day. We use close prices adjusted for splits and dividends when computing forward-looking returns.

We compare the performance of our portfolios built with various clustering algorithms with a portfolio built with industry classification data. The industry classification information maps each firm's SIC code to a single, non-overlapping Fama–French 12 industries sector label. The industries are nondurables (1), durables (2), manufacturing (3), energy (4),

chemicals (5), business equipment (6), telecommunications (7), utilities (8), shops (9), healthcare (10), finance (11), and other (12).

6.3.2 Performance of Portfolios

In this subsection, we investigate if the performance of the portfolios constructed with clustering algorithms is economically significant. The number of days used to estimate the number of clusters is 20 days, and the rolling window we use to estimate β and to compute the market residual return is 60 days. We set $w = 5$ days for the number of lookback days to construct the correlation matrix and to compute the cluster mean returns; the rebalance period to re-compute the correlation matrix, re-compute the clusters, and re-balance the portfolios is $\ell = 3$ days. The threshold to identify whether a stock is a previous winner or is a previous loser is $p = 0$, and we set the threshold $q = 5\%$ to consider that the portfolio mean-reverted. When we do not dynamically change the number of clusters, we fix the number of clusters to $K = 30$.

Table 6.1: Performances of statistical arbitrage portfolios with various clustering algorithms

	MP			90% Eigen			Fixed K		
Model	AR	SR	ST	AR	SR	ST	AR	SR	ST
SPONGE	10.99	1.02	1.81	11.90	1.07	1.89	10.21	1.01	1.80
SPONGE$_{sym}$	**12.05**	**1.11**	**2.01**	**12.20**	**1.10**	**2.01**	10.40	**1.03**	**1.80**
Spec	10.96	1.03	1.82	10.84	0.98	1.75	10.03	0.99	1.72
Lap$_{sym}$	11.19	0.91	1.60	11.24	0.88	1.55	**11.10**	0.97	1.66
Lap$_{rw}$	10.38	0.85	1.47	11.26	0.90	1.56	10.95	0.96	1.64
FF12	-	-	-	-	-	-	10.13	1.08	1.90
SPY	-	-	-	-	-	-	6.59	0.32	0.50

Table 6.1 presents the performance of the statistical arbitrage portfolios constructed with various clustering algorithms. The tabs "MP", "90% Eigen", and "Fixed K" represent methods for dynamically determining the number of clusters. For "MP", we use the

number of eigenvalues that exceed the upper boundary λ^+ of the Marchenko–Pastur distribution. This evaluation is performed on the correlation matrix derived from the returns matrix of dimension T by N. For "90% Eigen", the number of largest eigenvectors of the correlation matrix required to account for 90% of the total variance is the number of clusters we extract. For "Fixed K", we compute 30 clusters. In the table, "AR" is annualized return, "SR" is Sharpe ratio, and "ST" is Sortino ratio.

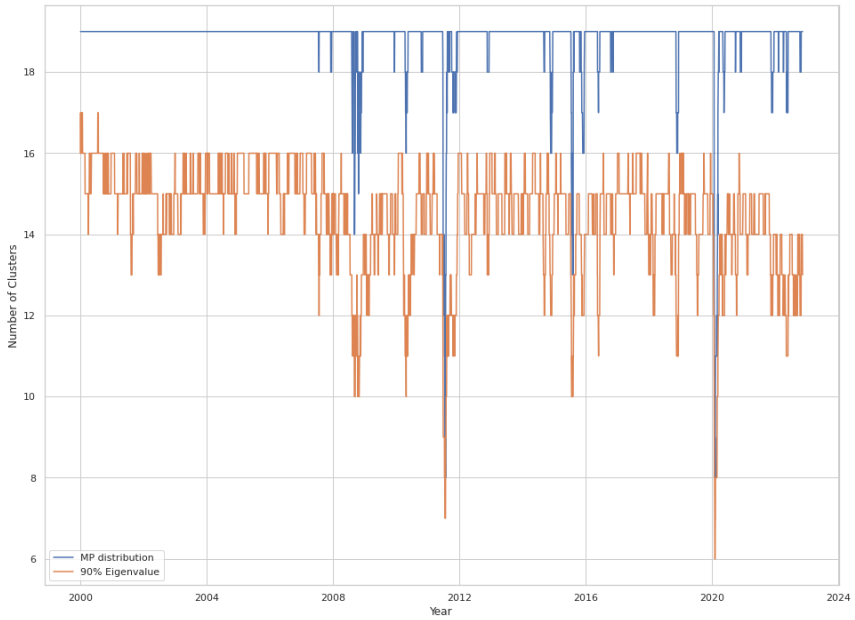


Figure 6.2: Historical number of clusters chosen by the Marchenko–Pastur distribution and total variance explained methods.

Figure 6.2 presents the number of clusters chosen by various methods. For "MP distribution", we use the number of eigenvalues that exceed the upper boundary λ^+ of the Marchenko–Pastur distribution. For "90% Eigen", we use the number of largest eigenvectors of the correlation matrix required to account for 90% of the total variance. Both methods provide relatively stable number of clusters, but both methods undergo drops in the number of clusters they find during financial hardships of the United States. For example, both methods experience a large drop in number of clusters they find during the 2008 financial crisis, August 2011 when U.S. credit rating was downgraded for the

first time in history, and COVID in 2020. This observation shows that the methods that dynamically determine the number of clusters can capture changes in market dynamics, especially when there is significant downside risks in the market.

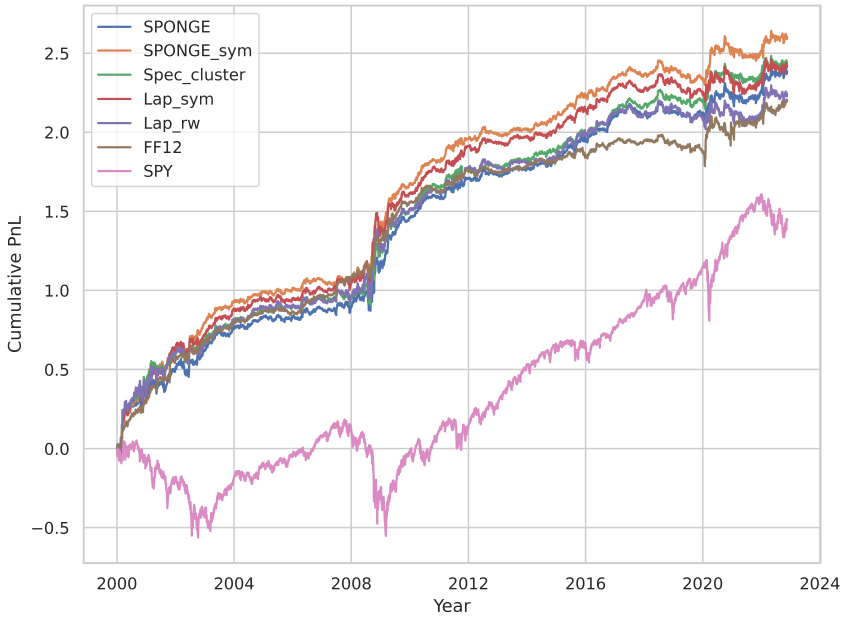


Figure 6.3: Cumulative returns of various strategies. The number of clusters is determined by the Marchenko–Pastur distribution. The cumulative returns are the sum of the daily returns without compounding.

The annualized returns of all portfolios are higher than 10%, where the $SPONGE_{sym}$ clustering portfolio delivers the highest overall performance in terms of annualized return, Sharpe ratio, and Sortino ratio. In particular, for the two methods where we dynamically determine the number of clusters, the $SPONGE_{sym}$ clustering portfolio has a Sortino ratio of 2.01. Overall, Our portfolios have similar Sharpe ratio and Sortino ratio as that of the Fama–French benchmark portfolio that uses the Fama–French 12 sectors.

We test the statistical significance of the Sharpe ratios obtained by the strategies, see Bailey and López de Prado (2014). Specifically, the Sharpe ratios of all arbitrage strategies are statistically significant at the 0.01% confidence level, while the Sharpe ratio of the SPY is not statistically significant at the 10% confidence level.

The performance of strategies in Table 6.1 is similar across various choices of number of clusters; thus, choosing the number of clusters dynamically does not significantly change the performance of portfolios. This observation lends strong support to the idea that the construction of arbitrage portfolios within each cluster is robust to the number of clusters.

Figure 6.3 shows the cumulative sum of returns of various strategies. In the long term, the statistical arbitrage portfolios tend to perform better when the volatility of the SPY portfolio is large (e.g., during the 2008 financial crisis). The performance of the Fama–French 12 sector portfolio is similar to that of the clustering-driven portfolios up until 2008, and the returns of the clustering-driven portfolios are higher than that of the Fama–French 12 sector portfolio after 2008.

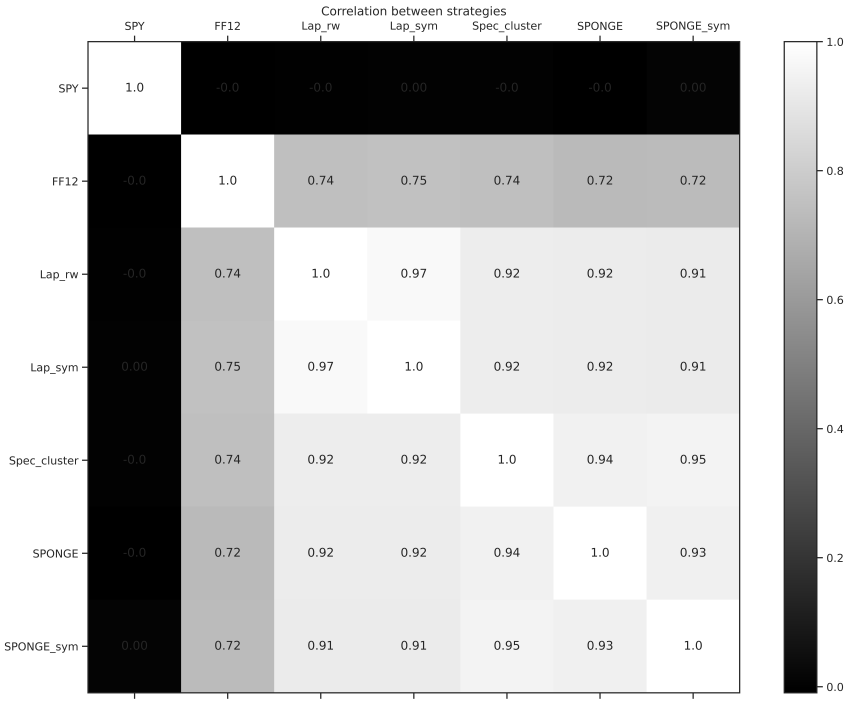


Figure 6.4: Correlation between returns of various strategies. The number of clusters of the clustering portfolios is determined by the Marchenko–Pastur distribution. Correlation coefficients are the Pearson correlation between the returns of strategies.

Figure 6.4 shows the correlation between the returns of the strategies we use to construct the portfolios. The correlation between all statistical arbitrage portfolios and the SPY

is close to zero, which confirms that the arbitrage portfolios are market neutral. When the clusters are computed with data-driven algorithms, the correlations among strategies are very high; this illustrates that the mean-reverting patterns that the clusters detect are similar. On the other hand, the correlation between the data-driven clustering portfolios and the Fama–French sector portfolio is much lower. The lower correlations with the Fama–French sector portfolio show that the performances of the clustering arbitrage portfolios cannot be fully explained by the underlying intra-sector relationships of the securities.

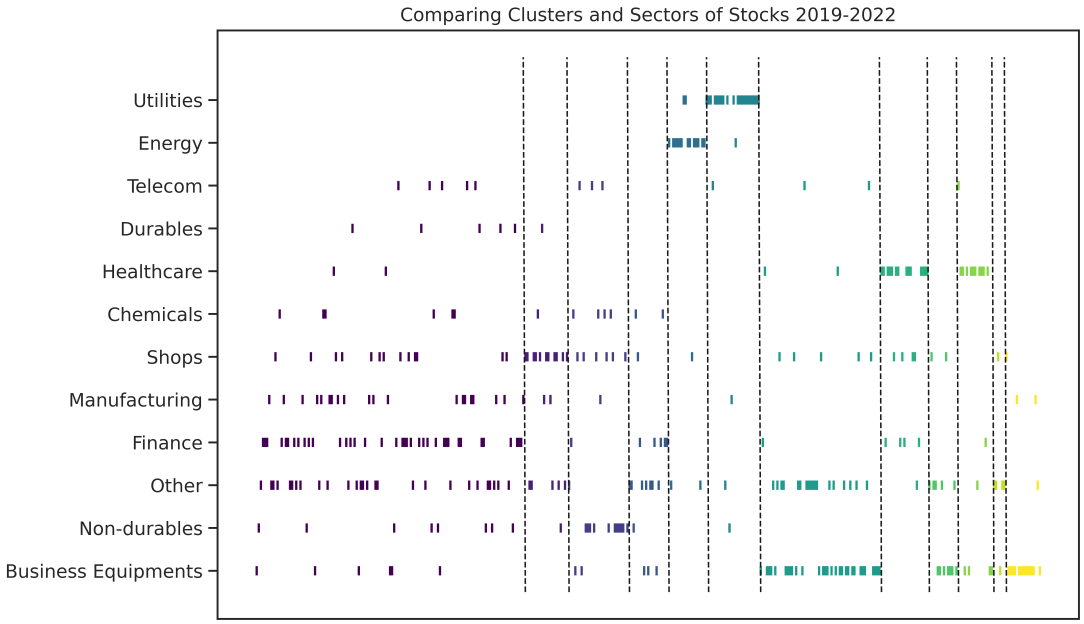


Figure 6.5: Comparison between the clusters created with the SPONGE algorithm on the correlation matrix of stocks to detect 12 clusters and the underlying Fama–French sector labels from 1 January 2019 to 31 December 2022. The area between black vertical dashes represents each cluster formed with SPONGE clustering. There are 377 stocks that are traded every day in this time period.

Figure 6.5 compares the clusters detected with SPONGE clustering and the stocks of the underlying Fama–French 12 industries. From 2019 to 2022, there are 377 stocks that are traded on every trading day. The SPONGE algorithm detects clusters that have large overlap with the following sectors: Utilities, Energy, Business Equipment, and Healthcare.

Other clusters detected by the SPONGE algorithm do not show strong alignment with any particular sectors; in particular, there is a large cluster that the SPONGE algorithm detects which contains stocks from all sectors except for Utilities and Energy. This observation further supports that the performance of the clustering statistical arbitrage portfolios cannot be fully explained by intra-industry mean-reversion behavior.

Table 6.2: Adjusted Rand Index between Clusters from various algorithms and the Fama–French 12 Sector labels from 1 January 2019 to 31 December 2022.

	ARI(%)
SPONGE	14.9
SPONGE_{sym}	13.5
Spectral	15.2
Laplacian_{rw}	14.5
Laplacian_{sym}	13.2

Finally, we use the Adjusted Rand Index (ARI) to measure the similarity between the clusters we detect using clustering algorithms and the Fama–French 12 sector labels, see Table 6.2. The ARI between the clusters of the algorithms and the Fama–French 12 sector labels is low. The similarity between clusters found by the clustering algorithms and the sector labels is less than 15%. This observation supports the observation that the returns of our statistical arbitrage portfolios cannot be fully explained by industry memberships. Our portfolio discovers new mean-reversion patterns among various partitions of stocks.

6.4 Conclusion

In this paper, we presented a novel framework for constructing statistical arbitrage portfolios that uses state-of-the-art graph clustering algorithms. Our empirical results demonstrated that our approach generates economically significant, profitable portfolios. In our study with historical data, we also showed that our framework is robust to the choice of

number of clusters and the choice of clustering algorithms. Our study fills a gap in the literature by exploring the potential of clustering methods to create profitable statistical arbitrage strategies and by applying signed clustering algorithms to financial time series analysis. Our framework serves as a new evaluation criterion to assess if a clustering algorithm can accurately group stocks into clusters of similar returns.

Several graph clustering algorithms were introduced in the last decades, but the number of downstream tasks which employs the recovered clusters is limited, especially in a financial context. Our work opens further lines of investigation. For example, one can explore clustering and statistical arbitrage for cross-asset correlation matrices, or one can instead use other matrices as inputs to our framework (e.g., matrix of co-movement upon reacting to events). Exploring a variant of our framework on higher-frequency data (e.g., intraday minutely returns) could potentially also lead to interesting findings and profitable trading strategies with significant economic benefits.

7 | Alpha in Analysts

7.1 Introduction

Sell-side analyst research is a cornerstone of equity markets, serving as a key source of information for institutional and retail investors alike. For long, the academic literature has scrutinized sell-side analyst predictions and recommendations, where many studies examine how forecast errors, bias, and revisions contain valuable information for equity pricing (see, e.g., Womack (1996), Clement (1999)). Yet, despite a voluminous literature on earnings forecasts and recommendation revisions, relatively few papers study *price targets*-the explicit estimates of future stock prices (see, e.g., Brav and Lehavy (2003), Asquith et al. (2005)). Price targets offer a direct translation of an analyst's prediction into an implied 12-month stock return forecast (see Bradshaw et al. (2012)). We use the implied returns to form portfolios for each analyst, where the stock weights depend on the magnitude and sign of the analyst's return forecast.

Despite the growing attention on price targets, most prior work analyzes them at the stock level-investigating how different analysts covering the same firm arrive at disparate target prices (Brav and Lehavy (2003); Asquith et al. (2005); Bradshaw et al. (2012); Palley et al. (2024); Bonini et al. (2007)). In contrast, we view *each analyst as a portfolio manager* whose price targets are tradable signals, which we use to evaluate the performance of individual analysts. In this way, we evaluate the performance of these *analyst-implied portfolios*, and we gauge if analysts as *individuals* produce alpha beyond a naive buy-and-hold benchmark of the same stocks they cover.

In the first part of this paper, we construct a long-short strategy for each analyst: we

go long stocks with positive implied returns (target price above current price) and short those with negative implied returns, where the portfolio weights are proportional to the magnitude of the forecasted returns. Next, we measure each analyst's alpha against an equally-weighted long-only portfolio of the same stocks. Our empirical tests use a large, comprehensive dataset of analyst price targets from IBES and stock specific information from CRSP, spanning more than two decades. Consistent with some prior findings on analyst stock-picking efficacy (Mikhail et al. (1999); Bradshaw et al. (2012)), our results show that, on average, the typical analyst does not generate statistically significant alpha when regressed on a naive baseline. However, a *nontrivial subset* of analysts does exhibit substantial forecasting skill, with statistically significantly positive alpha over time.[1]

Building on this heterogeneity in performance, we develop a **fund-of-analysts** framework, wherein an investor allocates capital across analysts based on the forecasts of each analyst's performance. Conceptually, this parallels a "fund-of-funds" model: the investor forms a meta-portfolio of analyst-implied portfolios, where the weights are dynamically updated according to each analyst's track record. First, we use characteristics of analysts-such as historical performance, volatility of implied returns, and coverage breadth-to forecast the profitability of each analyst's implied portfolios. Then, based on our predictions of analyst's future returns, we build a meta-portfolio where we allocate more wealth to analysts with higher predicted returns and less wealth to analysts with lower predicted returns. Our results demonstrate that dynamically allocating wealth to each analysts' implied portfolios (i.e., tilting away from predicted underperformers and toward those who we believe will perform well) can yield statistically significant alpha when regressed on standard passive

[1] A relevant piece of work that studies analyst performance is Stickel (1992) who finds that the earnings forecasts from all-American analysts (i.e. star analysts) tend to be more accurate than those who are not all-American analysts.

(i.e., buy and hold) benchmarks.

Overall, our contribution is twofold. First, we introduce a novel approach to measuring *analyst-level* skill, considering the analyst as a portfolio manager. Second, we develop a "fund-of-analysts" approach to show that one can extract alpha from a large and diverse panel of analysts-a perspective that has received little attention in the empirical literature on the sell-side research. Our findings complement existing evidence on the informative value of analyst recommendations (Tamura (2002); Gleason and Lee (2003); Da and Schaumburg (2011); Bradshaw et al. (2012)) and extend it to price targets, and we emphasize that understanding analyst heterogeneity is critical for investors. Our results are also related to a strand of the literature that corrects the analysts' forecasts to produce better investment signals (see Dechow and You (2020); Loudis (2024)), where our contributions provide a new perspective which predicts analyst portfolio returns instead of correcting forecasts for each stock.

The remainder of the paper is structured as follows. Section 2 introduces the methodology to construct implied portfolios and outlines the regression framework to estimate alpha for each analyst. Section 3 describes our data sources from CRSP and IBES. Section 4 presents empirical results on individual analysts' alpha and implements the fund-of-analysts approach, highlighting how investors dynamically allocate capital across analysts to exploit temporal differences in forecasting skill. Section 5 concludes.

7.2 Setup

For each calendar month m, suppose analyst i covers stocks $s_1, \ldots, s_N$ with current prices $P_1^m, \ldots, P_N^m$ and provides 12-month price targets $\hat{P}_1^{m+12}, \ldots, \hat{P}_N^{m+12}$ for each stock. Then,

for each stock, the price target of analyst i implies the return forecast

$$\hat{R}_k^{m+12} = \frac{\hat{P}_k^{m+12} - P_k^m}{P_k^m},$$

where $\hat{R}_k^{m+12}$ denotes the implied return for asset k at 12 months after month m.

Now, suppose an investor trusts analyst i and invests in a long-short portfolio based on the price forecasts of the analyst. The investor takes long positions in stocks with positive return forecasts and takes short positions in stocks with negative return forecasts. Additionally, the weight of the investor's long-short strategy is proportional to the magnitude of the return forecasts implied by the analyst. To compute the weights, let

$$K_+ = \left\{ k : \hat{R}_k^{m+12} > 0 \right\} \quad \text{and} \quad K_- = \left\{ k : \hat{R}_k^{m+12} < 0 \right\}$$

denote the set of stocks with positive and negative forecasted returns, respectively. Then, each stock $k \in K_+$ (i.e., each stock with positive implied return) is assigned a long position with weight

$$w_k^+ = \frac{\hat{R}_k^{m+12}}{\sum_{j \in K_+} \hat{R}_j^{m+12}},$$

and each stock $k \in K_-$ (i.e., each stock with negative implied return) is assigned a short position with weight

$$w_k^- = \frac{-\hat{R}_k^{m+12}}{-\sum_{j \in K_-} \hat{R}_j^{m+12}}.$$

The absolute value of these weights sum to one in each of the long and short buckets,

respectively, whenever K_+ and K_- are nonempty. Therefore, the investor constructs a self-financing long-short portfolio based on the recommendations made by analyst i.[2]

Next, let R_k^{m+12} denote the *realized* return of stock k between month m and month $m+12$, that is,

$$R_k^{m+12} \;=\; \frac{P_k^{m+12} - P_k^m}{P_k^m}\,,$$

where P_k^{m+12} is the price of stock k observed 12 months after m.

Therefore, the profit and loss (PnL) of the long-short portfolio, following the recommendations of analyst i, for the investor's strategy at month m marked to month $m+12$ is

$$\mathrm{PnL}_i^{m+12} \;=\; \sum_{k\in K_+} w_k^+\, R_k^{m+12} \;-\; \sum_{k\in K_-} w_k^-\, R_k^{m+12}\,.$$

Here, PnL_i^{m+12} is also a measurement of the overall quality of analyst i's price targets. Overall, if the analyst's price targets forecasts are accurate, the PnL incurred from the long-short portfolio should be large and positive. In particular, if an analyst's recommendation is valuable, a regression of the alpha of the analyst's PnL against an equally-weighted portfolio taking long positions on all the stocks the analyst covers should be positive and statistically significant. Consider

$$\mathrm{M}_i^{m+12} = \frac{1}{n}\sum_{x=1}^{n} R_x^{m+12}\,,$$

[2]When either K_+ or K_- is empty, the portfolio is no-longer self-financing and the investor takes a net long or short position of stocks.

where M_i^{m+12} denotes the PnL from trading an equally weighted long-only portfolio on all the stocks that analyst i covers in month m, and run the regression

$$\mathrm{PnL}_i^{m+12} = \alpha_i + \beta\, \mathrm{M}_i^{m+12} + \epsilon_{m+12} \tag{7.1}$$

for each analyst i over their career span. If the α of an analyst is positive and statistically significant over their career, there is evidence to believe that they are skilled and that they provide valuable insights into predicting stock returns.[3]

Now, assume that an investor has full access to all the analysts price forecasts, and the investor uses a "fund-of-fund" strategy that invests her wealth into portfolios formed from each analyst's implied portfolio. That is, the investor considers each analyst as a portfolio manager, and she invests a portion of her wealth on each of the analyst's implied portfolios.

Let c_i^m denote the proportion of the investor's wealth allocated as advised by analyst i in month m, assuming a total of I analysts, then twelve months later the investor collects

$$\mathrm{PnL}^{m+12} = \sum_{i=1}^{I} c_i\, \mathrm{PnL}_i^{m+12}\,.$$

A special case, analogous to the concept of "consensus" in the literature, arises when the investor is agnostic to the ability of each analyst, so she chooses to believe in each of them equally. In that case, the investor collects a "consensus PnL"

$$\mathrm{PnL}_{m+12}^{\mathrm{Consensus}} = \frac{1}{I} \sum_{i=1}^{I} \mathrm{PnL}_i^{m+12}\,.$$

[3] An analogy of why we run the regression in Equation (7.1) is when an investor holds a collection of stocks, and an analyst provides some recommendation to hold or sell the stock. Regression (7.1) informs us how much additional value the recommendations of this particular analyst brings to the investor.

Similar to investing according to the forecasts of individual analysts, the investor also wants to understand if, collectively, the analysts provide useful information in addition to a strategy that buys and holds all stocks. Consider the universe of stocks $s_1, \ldots, s_N$ that are covered by at least one analyst; an uninformed investor can choose to invest equally in these stocks and collect

$$\mathrm{PnL}_{m+12}^{\mathrm{Uninformed}} = \frac{1}{N} \sum_{b=1}^{N} \mathrm{R}_b^{m+12}$$

in month $m + 12$. If the collection of analysts provides additional value to the investor, then the intercept α in the regression

$$\mathrm{PnL}^{m+12} = \alpha + \beta \, \mathrm{PnL}_{m+12}^{\mathrm{Uninformed}} + \epsilon_m$$

should be positive and statistically significant.

Consider an investor who estimates analyst skills and ranks the analysts from best to worst, and she uses this ranking to build a portfolio. One way to rank analysts is to model the PnL incurred by the analyst as a function of the individual characteristics of the analyst. In this way, each month, the investor makes a prediction $\widehat{\mathrm{PnL}}_i^{m+12}$ of each analyst's future PnL. Consider

$$\mathrm{PnL}_i^{m+12} = g(\mathrm{A}_i \mid \mathrm{F}_m) \,,$$

where the PnL of analyst i from month m to month $m + 12$ is a function g of the analyst-specific characteristics A_i conditional on the information set F_m available at time up until month m. Here, analyst-specific characteristics A_i may contain predictors such as

the previous performance of the analyst, how well the analyst ranks among their peers throughout their career, etc.

The prediction $\widehat{\mathrm{PnL}}_i^{m+12}$ is a proxy for how much the investor trusts each analyst. Therefore, the proportion of wealth c_i^m that the investor assigns to the analyst is proportional to $\widehat{\mathrm{PnL}}_i^{m+12}$. In particular, to make the long-short portfolio self-financing, the investor *longs* the implied portfolio of analysts whose predicted $\widehat{\mathrm{PnL}}_i^{m+12}$ is positive, and *shorts* the implied portfolio of those whose predicted $\widehat{\mathrm{PnL}}_i^{m+12}$ is negative, while ensuring that the total long allocation equals the total short allocation, i.e., the investor's strategy is self-financing.[4] Concretely, define

$$I_+ \;=\; \big\{\, i : \widehat{\mathrm{PnL}}_i^{m+12} > 0 \big\} \quad \text{and} \quad I_- \;=\; \big\{\, i : \widehat{\mathrm{PnL}}_i^{m+12} < 0 \big\}.$$

Then, for each analyst $i \in I_+$, the investor's "long weight" is

$$c_i^+ \;=\; \frac{\widehat{\mathrm{PnL}}_i^{m+12}}{\displaystyle\sum_{j \in I_+} \widehat{\mathrm{PnL}}_j^{m+12}},$$

and for each analyst $i \in I_-$, the investor's "short weight" is

$$c_i^- \;=\; \frac{-\,\widehat{\mathrm{PnL}}_i^{m+12}}{\displaystyle-\sum_{j \in I_-} \widehat{\mathrm{PnL}}_j^{m+12}},$$

so, by construction, $\sum_{i \in I_+} c_i^+ = 1$ and $\sum_{i \in I_-} c_i^- = 1$.

Because going "long" a positive forecast and "short" a negative forecast in each analyst's implied portfolio does not require net initial capital, the fund-of-analysts portfolio is self-

[4] The analysts' PnLs are PnL of long-short portfolios themselves, so we can short an analyst's implied portfolio.

financing. Then, the total realized *informed* PnL at month $m + 12$ becomes

$$\mathrm{PnL}_{m+12}^{\mathrm{Informed}} = \sum_{i \in I_+} c_i^+ \, \mathrm{PnL}_i^{m+12} - \sum_{i \in I_-} c_i^- \, \mathrm{PnL}_i^{m+12},$$

where PnL_i^{m+12} is the *realized* PnL each analyst i generates from their implied portfolio over the same 12-month horizon. Thus, the investor, informed by forecasts of analyst performances, constructs a "fund-of-analysts", allocating larger absolute weights to the analysts with higher predicted performance (for longs) or with lower predicted performance (for shorts).

Similarly, we test if this "fund-of-analysts" approach results in statistically significant alpha when regressed on a long-only, equally-weighted strategy that buys all stocks covered by at least one analyst. In this way, we test if the price forecasts for sell-side investment analysts contain useful information in predicting stock returns over the one-year horizon.

7.3 Data

Stock returns, firm characteristics, and other data are from the WRDS online database. We select all common stocks publicly traded on the NYSE (New York Stock Exchange), AMEX (American Stock Exchange), and NASDAQ (National Association of Securities Dealers Automated Quotations). Daily market data, such as prices, trading volumes, and returns, are from the Center for Research in Security Prices (CRSP). Analyst price forecasts are from IBES (Institutional Brokers' Estimate System).

The starting date for each analyst's price forecasts is the $ACTDATS$ column in the IBES database, and we use data from CRSP to compute the 12-month implied return. Our

sample contains all analyst recommendations available from January 1999 to November 2024.

7.4 Results

First, we examine if individual stock analysts, when viewed as portfolios, carry statistically significant alpha against an equally-weighted portfolio of the stocks they cover.

Table 7.1: Performance metrics by analyst experience

Performance metrics for analysts categorized by years of experience. The metrics include α, the abnormal return estimated from a factor model against equally weighted portfolio of stocks the analyst covers, with corresponding cross-sectional t-statistics in parentheses. The parameter β represents the market beta, capturing systematic risk exposure. Return denotes the average annualized return of the analysts' stock recommendations, with t-statistics shown in parentheses. Baseline is the average return of the benchmark portfolio, which is an equally-weighted long-only portfolio holding the stocks that the analyst covers. Stocks represents the average number of stocks covered per analyst in each experience group in each month.* Count refers to the total number of analysts in each category. Sig. 5% and Sig. 1% indicate the number of analysts whose α estimates are statistically significant at the 5% and 1% levels, respectively.

Experience	α	β	Return	Baseline	Stocks	Count	Sig. 5%	Sig. 1%
<1 year	0.002	0.766	0.020	0.025	1.64	3491	182	61
	(0.30)		(2.02)					
1-2 years	0.001	0.792	0.026	0.046	2.11	1920	143	40
	(0.30)		(2.68)					
2-5 years	−0.001	0.754	0.054	0.084	2.67	2169	256	89
	(−0.19)		(7.72)					
5+ years	0.001	0.726	0.079	0.114	3.68	2242	281	107
	(0.61)		(18.25)					
Overall	0.001	0.758	0.042	0.062	2.43	9822	862	297
	(0.41)		(9.60)					

* Not all analysts release forecasts for every stock they cover each month. E.g., on average an analyst covers 5.6 stocks each year, but they report forecasts for 2.4 stocks each month. Over the lifetime of an analyst, they cover, on average, 13.3 stocks.

Table 7.1 shows that investment analysts, on average, exhibit statistically insignificant α regressed on an equally-weighted portfolio of stocks that they cover. The α from price

targets of analysts are small in magnitude and not significantly different from zero across all experience buckets of analysts. For analysts at different experience levels and for all analysts on average, their price targets imply a portfolio with statistically significant returns that generally increase with their experience level. In particular, analysts with more than five years of experience exhibit the highest raw return of approximately 8% for each of the 12-months portfolio the analysts hold each month. However, for each bucket of analyst experiences, the portfolio implied by the price targets of analysts cannot outperform the equally-weighted long-only portfolio with the stocks each analyst covers. On average, the baseline outperforms the analyst implied portfolios by around 2% per month.

In general, analysts with a longer career span tend to follow more stocks, and the stocks that the analysts follow tend to be more successful (the average return of the baseline is higher for analysts who survive longer). Thus, we ask if analysts survive longer because they cover more successful stocks by chance or because they pick stocks with more growth potential. To explore this question, we count the number of analysts with different career lengths who exhibit statistically significant α at 5% and 1% significance levels.

Recall that, although we find that the portfolios implied from analysts' price targets exhibit statistically insignificant α on average, individual analysts can still demonstrate significant outperformance over simple buy and hold portfolios. Specifically, if the number of analysts with significant α exceeds what would be expected under a random binomial distribution, this would suggest that at least some analysts may possess genuine stock price forecast skill rather than benefiting purely from chance.

The results in Table 7.2 indicate that the percentage of analysts with statistically signif-

Table 7.2: Fraction of analysts with significant α and binomial tests by experience

For each experience group: total number of analysts (Count), number of analysts with statistically significant α at the 5% level (Sig. 5%) and 1% level (Sig. 1%), fraction of the group these analysts represent (% Sig.), and one-sided binomial test p-value. The binomial test null hypothesizes that the true fraction of analysts with significant α is exactly 5% (or 1%), i.e., arising purely by chance. A low p-value indicates that more analysts are significant than would be expected if there were no skill component in their forecasts.

Experience	**Count**	**Significant at 5%**			**Significant at 1%**		
		Number	**% Sig.**	***p*-value**	**Number**	**% Sig.**	***p*-value**
<1 year	3,491	182	5.21%	0.2919	61	1.74%	0.0000
1–2 years	1,920	143	7.45%	0.0000	40	2.08%	0.0000
2–5 years	2,169	256	11.81%	0.0000	89	4.10%	0.0000
>5 years	2,242	281	12.53%	0.0000	107	4.77%	0.0000
Overall	9,822	862	8.78%	0.0000	297	3.02%	0.0000

icant α generally increases with experience. Among analysts with less than one year of experience, 5.21% achieve significance at the 5% level, a proportion that is not significantly different from what would be expected under the null ($p = 0.284$). However, at the 1% level, 1.74% of analysts exceed the threshold, which is significantly higher than the expected 1% rate. For analysts with 1–2 years of experience, 7.45% of them achieve significance at the 5% level, suggesting that a higher-than-random fraction of these analysts exhibits genuine forecasting ability. Similarly, at the 1% level, 2.08% of analysts exceed the threshold, reinforcing the notion that skill is emerging within this group. Among more experienced analysts, the fraction with statistically significant α rises sharply. For those with 2–5 years of experience, 11.81% are significant at the 5% level , and 4.10% at the 1% level. The trend continues for the most experienced analysts (more than 5 years), where 12.53% of them are significant at the 5% level and 4.77% at the 1% level, both highly significant results.

Across all analysts, 8.78% demonstrate significant α at the 5% level, with a corresponding

binomial test $p = 0.000$, strongly rejecting the null hypothesis that skill is entirely absent. Similarly, 3.02% are significant at the 1% level, further reinforcing the idea that some analysts do exhibit genuine forecasting ability.

The results in Table 7.2 suggest that while analysts, on average, do not generate statistically significant α, the proportion of those who do increases with experience. This provides preliminary evidence that while it is possible that some analysts survive longer because they cover inherently better-performing stocks, at least some analysts do develop skill in forecasting stock returns.

Recall that Table 7.1 shows that the baseline return-the average return of an equally-weighted portfolio of covered stocks-rises with experience, indicating that more seasoned analysts tend to follow stronger-performing stocks. This suggests that luck in stock assignment is a factor in survival of analysts. Next, we examine if the proportion of analysts with significant α also increases with tenure. Table 7.2 shows that the proportion of analysts with significant α at both the 5% and 1% levels far exceeds what we expect from random assignments, particularly among more experienced analysts. This suggests that, beyond merely covering better stocks, at least some analysts improve their ability to identify potential trading opportunities over time.

Now, consider investors who have access to all analyst price targets. Recall we assume investors follow a "fund-of-analysts" strategy where proportions of their wealth invested in portfolios depends on how well they believe in each analyst's price targets.

First, we consider a naive strategy where the investors allocate equal proportions of their wealth into portfolios implied by each analyst's price targets. We compare the performance of this strategy to that of a naive baseline where the investor invests in an equally-weighted

long-only portfolio of all stocks that are covered by at least one analyst in the month of the forecast (i.e., the union of all stocks covered by analysts). We also consider an infeasible, perfect foresight strategy where the investor knows beforehand which analysts are in the top or bottom 2.5% among their cohort in that month. We include these infeasible strategies to motivate that an investor might want to choose to believe in some analysts more or less than in others and trade a "fund-of-analysts" strategy.

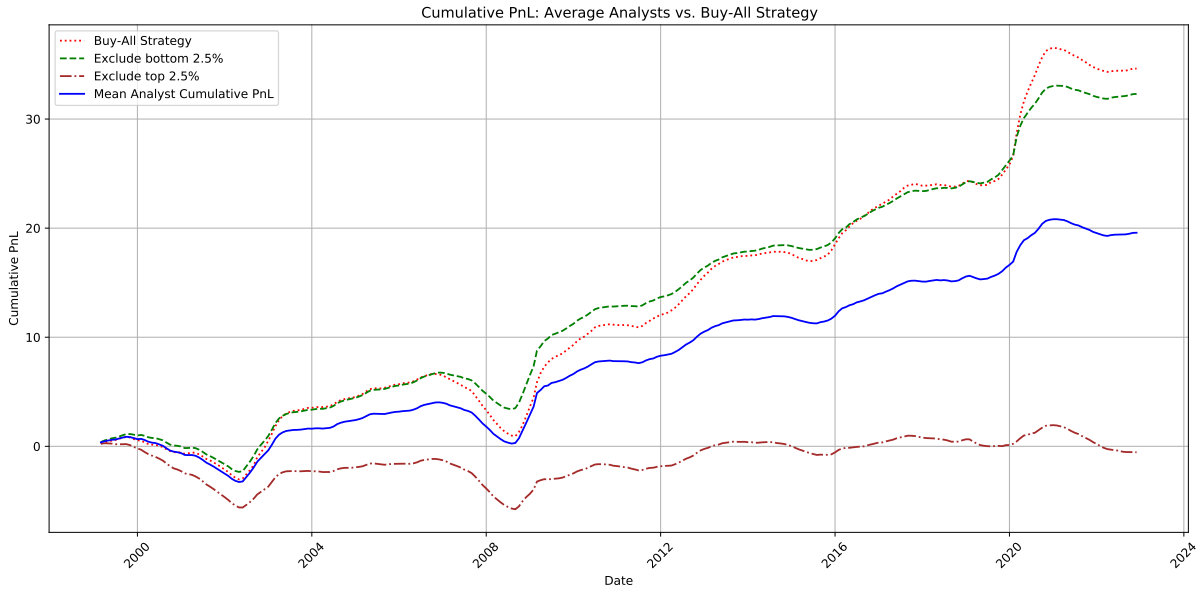


Figure 7.1: Cumulative returns of fund-of-analysts strategies.

Figure 7.1 plots the cumulative returns of the four portfolio strategies evaluated in Table 7.1. The Baseline strategy (red dash-dots) is the passive benchmark, while the Mean Analyst strategy (blue solid line) represents an equally-weighted allocation across analysts. The Exclude Bottom 2.5% strategy (green dashes) removes the worst-performing analysts, while the Exclude Top 2.5% strategy (brown dots) excludes the best-performing analysts, where both assume perfect foresight.

In Figure 7.1, the naive strategy of buying all the analysts underperforms a buy-and-hold strategy with all stocks, showing that an average analyst does not have any stock price forecasting skill. However, the performance of the perfect foresight infeasible strategies,

where the investor knows beforehand who the best and worst analysts are, show that there is considerable heterogeneity in forecasting performance of analysts in any given month. Also, if the investor excludes the worst analysts and invests equally in all other analysts, the performance of her portfolio would be similar to that of being long the market. On the other hand, note that the returns of the portfolio that excludes the bottom 2.5% analysts is nearly zero. This shows that on any given month, only a small number of analysts provides useful price targets while others can be roughly perceived as providing only noise to investors.

Table 7.3: Performance metrics of fund-of-analysts strategies

Performance metrics for different fund-of-analysts strategies. Mean Analyst refers to the strategy that takes equally weighted positions in all analyst portfolios; Exclude Bottom 2.5% and Exclude top 2.5% refer to the perfect foresight infeasible strategy where the investor knows which analysts will perform poorly/well at the 2.5% percentile; Baseline refers to a strategy where the investor takes long-only equally weighted portfolios in all stocks that analysts cover. Here, Mean Return is the average monthly return of the strategy; Volatility is the standard deviation of monthly returns; Sharpe is the mean return divided by volatility, assuming a risk-free rate of zero; α represents the abnormal return estimated from a regression against the Baseline, with the corresponding t-Stat testing its significance; and β represents the market beta, capturing systematic risk exposure relative to the Baseline.

Strategy	Mean Return	Volatility	Sharpe	Alpha	t-Stat	Beta
Mean Analyst	0.0685	0.2004	0.3416	−0.0034	−0.674	0.5932
Exclude Bottom 2.5%	0.1130	0.2407	0.4693	0.0226	5.528	0.7466
Exclude Top 2.5%	−0.0019	0.1538	−0.0124	−0.0523	−9.919	0.4166
Baseline	0.1211	0.3107	0.3896	–	–	–

In Table 7.3, we present the values of the performances of strategies in Figure 7.1. The Mean Analyst strategy, which weights the portfolios derived from every analyst's price targets equally, realizes a monthly return of 0.0685 with 0.2004 volatility. Although the strategy's Sharpe ratio (0.3416) is positive, the α of the portfolio is statistically insignificant and marginally negative with a t-stat of -0.674 , underscoring the absence of the

outperformance relative to the Baseline. Consistent with Figure 7.1, the average analyst does not seem to provide any additional information to the market.

On the other hand, the Exclude Bottom 2.5% portfolio demonstrates that removing a small cohort of poor forecasters enhances risk-adjusted performance. After discarding these worst-performing analysts, with perfect foresight, the portfolio's mean return (0.1130) is close to that of the Baseline (0.1211), while its Sharpe ratio (0.4693) exceeds that of the Baseline (0.3896). Notably, the strategy achieves a statistically significant α of 0.0226 ($t = 5.528$), indicating that cutting out low-skill forecasters generates significant excess returns, even after accounting for systematic exposure ($\beta = 0.7466$).

Meanwhile, excluding the Top 2.5% performing analysts results in a substantial performance drag: the mean return becomes slightly negative (-0.0019) and the α is both large in magnitude (-0.0523) and highly significant ($t = -9.919$). This underscores that a small contingent of analysts (which changes over time) contributes disproportionately to overall stock price forecasting abilities.

Taken together, these findings highlight two key points. First, the average analyst does not reliably add value beyond a equally-weighted portfolio of the stocks they cover. Second, there exists a nontrivial subset of analysts whose forecasts do exhibit meaningful skill. The central challenge for investors is to identify these outperformers-or at least to filter out the worst performers-on a dynamic and systematic basis.

Motivated by this evidence, we next assume that an investor *models* how much credence to place in each analyst's forecasts, rather than naively pooling them all or discarding some of them indiscriminately. Specifically, the investor predicts each analyst's expected return at the beginning of each month. She updates her projections based on performance

signals, and she scales her investment weight in each analyst's implied portfolio according to these evolving beliefs. Specifically, when an analyst is predicted to have larger PnL, she invests in a larger proportion of her wealth into the analyst. By continually learning and re-weighting analysts' price targets, investors can pursue a more refined "fund-of-analysts" strategy that captures real forecasting skill. We use the refined "fund-of-analysts" strategy to re-evaluate if investors can extract meaningful information from analyst price targets.

The investor uses the past performance metrics of each analyst to model the future expected returns of portfolios based on each analyst's predictions.

We use an expanding regression approach to estimate each analyst's expected performance every month with data available up to 12 months before the given month. Specifically, for month t, we regress the realized PnL of each analyst's price targets up through $t-12$ months on a set of predictor variables that captures the analyst's recent accuracy, risk profile, and coverage breadth. With this restriction on the training window, we study a realistic investment scenario where month t's portfolio decisions use information available at the time of investment decisions, thus preventing any lookahead bias in our out-of-sample evaluation. We use all available previous data to retrain the regression model every month.

Note that we do not aim to produce the best possible analyst PnL prediction model. We include a class of intuitive predictors to mimic an environment where investors use analysts' previous performances to decide how much to believe in each analyst.

Table 7.4 provides an overview of the predictor variables we use in our rolling regression. These predictors capture key aspects of an analyst's historical forecasting performance. With a mix of ranking-based cross-sectional features and rolling historical measures, an

Table 7.4: Predictor variables and definitions

Predictor	**Definition**
performance percentile	Analyst's percentile rank by historical performance (accuracy, returns) relative to peers. Higher values indicate stronger track records.
stock percentile	Percentile rank of the breadth of an analyst's stock coverage, capturing how many tickers the analyst covers compared to peers.
recent percentile	Percentile rank of the analyst's most recent six-month rolling performance, highlighting recency effects in forecasting skill.
std percentile	Percentile rank of the analyst's return volatility, assessing relative riskiness of the analyst's calls.
rolling std 6m	The analyst's realized return volatility measured over a rolling six-month window up until 12 months before.
rolling mean 6m	The analyst's realized average return over a rolling six-month window up until 12 months before

investor captures both static analyst characteristics and dynamic changes in the analysts' forecasting effectiveness.

Once the regression estimates are obtained, we use the fitted model to predict each analyst's expected return from month t to month $t + 12$. Recall that we allocate positive (long) weights, proportional to the magnitude of their predicted PnL, to analysts forecasted to deliver positive PnL, whereas those projected to lose money receive negative (short) weights. This approach aligns with investors who adjust their decisions at every rebalance period, and the portfolio composition changes every month as new data arrive and as the relative performances of each analyst changes over time.

Figure 7.2 presents the cumulative PnL of the fund-of-analysts strategy after we predict analysts' future PnL and use the expanding-window regression to update the weight of each analyst in the portfolio. After fitting an expanding regression with six predictors, we achieve a portfolio that outperforms the baseline (and the average analyst) in terms of having higher Sharpe ratio and statistically significant alpha compared to the baseline.

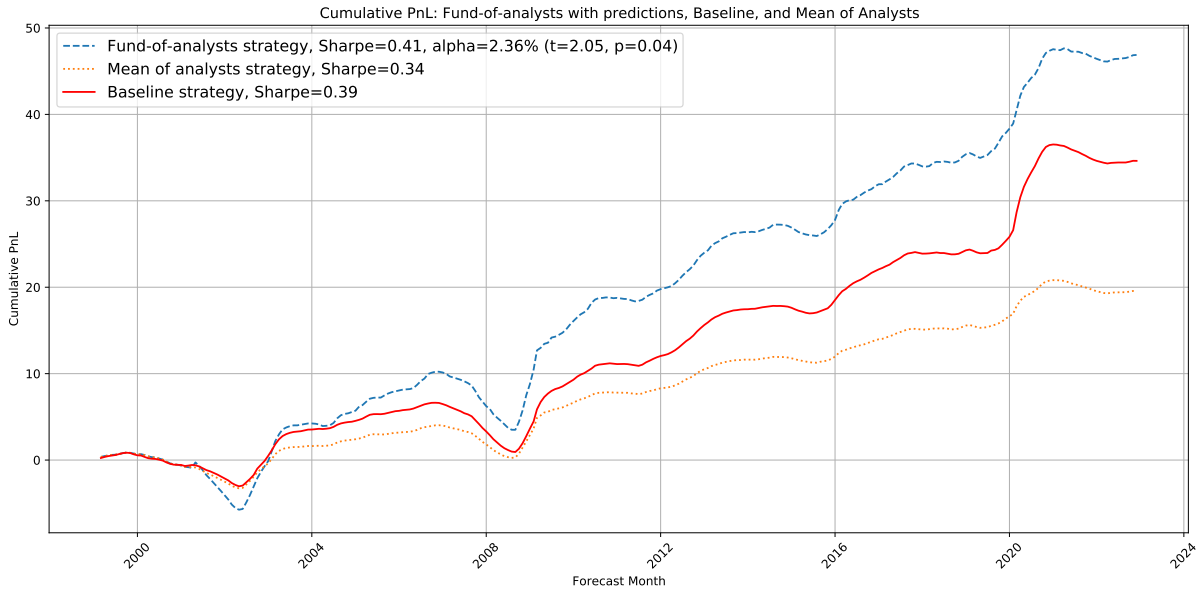


Figure 7.2: Cumulative returns of refined fund-of-analysts strategies.

In particular, the alpha of the strategy that uses our forecast of analysts price target performance is 2.36% per month and statistically significant when regressed on the baseline strategy of an equally weighted portfolio of all stocks that are covered by at least one analyst. This preliminary evidence shows that when appropriately adjusting for and selection of analysts' abilities among their cohort, investors can extract useful information from their price target forecasts beyond a simple buy-and-hold strategy.

Recall that we run the predictive expanding-window regression with the predictors in Table 7.4 every month. Next, we report the coefficient from the last regression window in Table 7.5.

From the perspective of a predictive regression model, we observe in Table 7.5 that analyst price target performance is positively correlated to their long-term cross-sectional ranking in performance and standard deviation, and the performance is negatively correlated to the number of stocks they cover, their short-term cross-sectional ranking in performance, and their short-term absolute performance. All these effects are strongly statistically

Table 7.5: Predictive Regression Results

Coefficients from the final expanding window regression in December 2023. The dependent variable is the one-year-ahead PnL of each analyst. The independent variables include several predictors related to performance, recent trends, and statistical properties listed in Table 7.4. The regression is estimated using ordinary least squares, with standard errors reported. The table includes significance tests for each coefficient.

Predictor	Coefficient	Std. Err	T-Stat	P-Value	2.5%	97.5%
performance_percentile	0.0469	0.013	3.706	0.000	0.022	0.072
recent_percentile	−0.0309	0.011	−2.889	0.004	−0.052	−0.010
rolling_std_6m	0.0193	0.002	8.489	0.000	0.015	0.024
rolling_mean_6m	−0.0216	0.004	−5.017	0.000	−0.030	−0.013
std_percentile	0.0580	0.011	5.356	0.000	0.037	0.079
stock_percentile	−0.0422	0.010	−4.268	0.000	−0.062	−0.023

significant at 1% level.

In other words, Table 7.5 shows that analyst price target forecasting ability exhibits long-term persistence and short term mean-reversion. Additionally, analysts who cover more stocks tend to perform worse.

Next, we explore how are our results related to exposures to common risk factors. We run a linear regression that explains the returns of our trading strategies using the Fama–French three factors. In particular, in each month, we take the 12-month forward-looking return of each factor and use these returns to explain the returns of our strategies.

Table 7.6 reports the alpha and exposures for the portfolios in Figure 7.2. We notice that after considering for the Fama–French 3 factors, the strategy where one first predicts analyst performances and then trades accordingly yields statistically significant monthly alpha whereas both the consensus analyst strategy and the long-only strategy yield statistically insignificant alpha. Additionally, after predicting analyst performances and trading accordingly, the portfolio gains a much larger exposure to the market and to the SMB factor, potentially suggesting that the trading strategy is more exposed to systematic risk

Table 7.6: Fama–French Three-Factor Regression Results

Results of regressing each strategy's excess returns on the Fama–French three-factor model. Alpha represents the intercept and measures the unexplained return of the strategy after controlling for systematic risk factors. The betas capture the exposure to the market (Mkt), size (SMB), and value (HML) factors. The t-statistics are reported in brackets. Statistical significance is denoted as follows: † $p < 0.05$, * $p < 0.1$, ‡ $p < 0.01$.

Strategy	Alpha	Mkt Beta	SMB Beta	HML Beta	R-squared
Regression Strategy	$0.0282^{\dagger}$	$1.5774^{\ddagger}$	$1.1420^{\ddagger}$	-0.0113	0.631
	(2.087)	(13.406)	(6.470)	(-0.110)	
Mean Strategy	-0.0022	$0.8205^{\ddagger}$	$0.5564^{\ddagger}$	-0.0468	0.689
	(-0.380)	(14.305)	(7.032)	(-1.004)	
Baseline Strategy	0.0024	$1.2910^{\ddagger}$	$1.2135^{\ddagger}$	0.0730	0.789
	(0.347)	(21.321)	(7.005)	(1.090)	

factors commonly associated with small-cap and market-wide movements. The significant market beta suggests that a substantial portion of the strategy's returns can be attributed to general market movements, while the high SMB beta indicates a tilt towards smaller firms.

The exposure to the HML factor is near zero and statistically insignificant for the Regression Strategy, implying that the portfolio does not inherently favor value or growth stocks. This is in contrast to the Baseline Strategy, which exhibits a slightly positive HML beta, albeit also statistically insignificant.

From a risk-adjusted return perspective, the significant positive alpha of the Regression Strategy suggests that after controlling for systematic risks, the strategy continues to generate excess returns. In contrast, both the Mean Strategy and the Baseline Strategy show alphas close to zero, indicating that their returns are likely explainable by systematic risk exposures rather than skill-based alpha generation.

7.5 Conclusion

In this paper, we revisited the role of sell-side analyst price targets, where we adopt a novel perspective that treats each analyst as a portfolio manager and use their price targets to construct 12-month return predictions. Our analysis shows that although the average analyst does not generate statistically significant alpha relative to a naive long-only benchmark, a nontrivial subset of analysts consistently exhibits genuine forecasting skill. This heterogeneity in performance suggests that both luck and skill play roles in analysts' long-term success.

Motivated by these findings, we introduced a fund-of-analysts framework that dynamically allocates capital across individual analyst-implied portfolios based on predictive signals derived from historical performance, volatility measures, and coverage breadth of each analyst. Our results demonstrate that underweighting poor forecasters and overweighting analysts with stronger track records can generate significant excess returns relative to passive strategies. In particular, our expanding-window regression approach, which updates investor beliefs each month, confirms that appropriately adjusting weights based on predicted performance enhances the risk-adjusted performance of the overall portfolio.

These insights extend the existing literature on the informational content of analyst price targets and offer practical guidance for investors seeking to extract value from sell-side research. Overall, our work provides a new perspective on analyst forecasting, emphasizing that, when harnessed correctly, the diversity in analyst views can be transformed into a meaningful investment signal.

Bibliography

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Akbas, F., Boehmer, E., Jiang, C., and Koch, P. D. (2022). Overnight returns, daytime reversals, and future stock returns. *Journal of Financial Economics*, 145(3):850–875.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets*, 5(1):31–56.

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249.

Arik, S. Ö. and Pfister, T. (2021). Tabnet: Attentive interpretable tabular learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6679–6687.

Asquith, A., Mikhail, S., and Au, J. (2005). Information content of equity analyst reports. *Journal of Financial Economics*, 78:229–263.

Avellaneda, M. and Lee, J. H. (2010). Statistical arbitrage in the US equities market. *Quantitative Finance*, 10(7):761–782.

Badrinath, S. G., Kale, J. R., and Noe, T. H. (1995). Of shepherds, sheep, and the cross-autocorrelations in equity returns. *The Review of Financial Studies*, 8(2):401–430.

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

Bansal, R. and Yaron, A. (2004). Risks for the long run: A potential resolution of asset pricing puzzles. *The Journal of Finance*, 59(4):1481–1509.

Barardehi, Y. H., Bogousslavsky, V., and Muravyev, D. (2022). What drives momentum and reversal? evidence from day and night signals. *SSRN Electronic Journal*.

Barber, B. M. and Odean, T. (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The Review of Financial Studies*, 21(2):785–818.

Basnarkov, L., Stojkoski, V., Utkovski, Z., and Kocarev, L. (2020). Lead–lag relationships in foreign exchange markets. *Physica A: Statistical Mechanics and its Applications*, 539:122986.

Bennett, S., Cucuringu, M., and Reinert, G. (2022). Lead-lag detection and network clustering for multivariate time series with an application to the us equity market.

Bergault, P., Drissi, F., and GuÃ©ant, O. (2022). Multi-asset optimal execution and statistical arbitrage strategies under ornstein-uhlenbeck dynamics.

Berkman, H., Koch, P. D., Tuttle, L., and Zhang, Y. J. (2012). Paying attention: Overnight returns and the hidden cost of buying at the open. *The Journal of Financial and Quantitative Analysis*, 47(4):715–741.

Bertram, W. K. (2010). Analytic solutions for optimal statistical arbitrage trading. *Physica A: Statistical Mechanics and its Applications*, 389(11):2234–2243.

Bessembinder, H. (2003). Trade execution costs and market quality after decimalization. *Journal of Financial and Quantitative Analysis*, 38(4):747–777.

Black, F., Jensen, M. C., and Scholes, M. (1972). The capital asset pricing model: Some empirical tests.

Blume, L., Easley, D., and O'Hara, M. (1994). Market statistics and technical analysis: The role of volume. *The Journal of Finance*, 49(1):153–181.

Bogousslavsky, V. (2021). The cross-section of intraday and overnight returns. *Journal of Financial Economics*, 141(1):172–194.

Bonini, S., Salvi, A., Bianchini, R., and Zanetti, L. (2007). Portfolio returns and target prices. *SSRN Electronic Journal*.

Boudoukh, J., Feldman, R., Kogan, S., and Richardson, M. (2019). Information, trading, and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3):992–1033.

Boyarchenko, N., Larsen, L. C., and Whelan, P. (2023). The overnight drift. *The Review of Financial Studies*, 36(9):3502–3547.

Bradshaw, M., Brown, L., and Huang, K. (2012). Do sell-side analysts exhibit differential target price forecasting ability? *Review of Accounting Studies*, 18.

Branch, B. S. and Ma, A. J. (2006). The overnight return, one more anomaly. *SSRN Electronic Journal*.

Brav, A. and Lehavy, R. (2003). An empirical analysis of analysts' target prices. *Journal of Finance*, 58:1163–1192.

Brennan, M. J., Jegadeesh, N., and Swaminathan, B. (1993). Investment Analysis and the Adjustment of Stock Prices to Common Information. *The Review of Financial Studies*, 6(4):799–824.

Buccheri, G., Corsi, F., and Peluso, S. (2019). High-frequency lead-lag effects and cross-asset linkages: A multi-asset lagged adjustment model. *Journal of Business & Economic Statistics*, 39:1–28.

Campbell, J. Y. and Cochrane, J. H. (1995). By force of habit: A consumption-based explanation of aggregate stock market behavior. Working Paper 4995, National Bureau of Economic Research.

Cartea, A., Gan, L., and Jaimungal, S. (2019). Trading co-integrated assets with price impact. *Mathematical Finance*, 29(2):542–567.

Cartea, A. and Jaimungal, S. (2016). Algorithmic trading of co-integrated assets. *International Journal of Theoretical and Applied Finance*, 19(06):1650038.

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.

Chevyrev, I. and Kormilitzin, A. (2016). A primer on the signature method in machine learning.

Chordia, T., Subrahmanyam, A., and Anshuman, V. R. (2001). Trading activity and expected stock returns. *Journal of Financial Economics*, 59(1):3–32.

Chordia, T. and Swaminathan, B. (2000). Trading volume and cross-autocorrelations in stock returns. *Journal of Finance*, 55:913–935.

Clement, M. B. (1999). Analyst forecast accuracy: Do ability, resources, and portfolio complexity matter? *Journal of Accounting & Economics*, 27:285–303.

Cooper, M. J., Cliff, M. T., and Gulen, H. (2008). Return differences between trading and non-trading hours: Like night and day. *SSRN Electronic Journal*.

Cucuringu, M. (2015). Sync-rank: Robust ranking, constrained ranking and rank aggregation via eigenvector and semidefinite programming synchronization.

Cucuringu, M., Davies, P., Glielmo, A., and Tyagi, H. (2019a). Sponge: A generalized eigenproblem for clustering signed networks.

Cucuringu, M., Li, H., Sun, H., and Zanetti, L. (2019b). Hermitian matrices for clustering directed graphs: insights and applications. *CoRR*, abs/1908.02096.

Cucuringu, M., Singh, A. V., Sulem, D., and Tyagi, H. (2021). Regularized spectral methods for clustering signed networks. *Journal of Machine Learning Research*, 22(264):1–79.

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

Da, Z. and Schaumburg, E. (2011). Relative valuation and analyst target price forecasts. *Journal of Financial Markets*, 14(1):161–192.

De Bacco, C., Larremore, D. B., and Moore, C. (2018). A physical model for efficient ranking in networks. *Science advances*, 4(7):eaar8260.

Dechow, P. M. and You, H. (2020). Understanding the determinants of analyst target price implied returns. *The Accounting Review*, 95(6):125–149.

DeMiguel, V., Gil-Bazo, J., Nogales, F. J., and Santos, A. A. (2023). Machine learning and fund characteristics help to select mutual funds with positive alpha. *Journal of Financial Economics*, 150(3):103737.

DeMiguel, V., Nogales, F. J., and Uppal, R. (2014). Stock Return Serial Dependence and

Out-of-Sample Portfolio Performance. *The Review of Financial Studies*, 27(4):1031–1073.

Dennis, P., Perfect, S., Snow, K., and Wiles, K. (1995). The effects of rebalancing on size and book-to-market ratio portfolio returns. *Financial Analysts Journal*, 51:47–57.

Doyle, J. and Magilke, M. (2008). The timing of earnings announcements: An examination of the strategic disclosure hypothesis. *Accounting Review*, 84.

Elliott, R. J., Hoek, J. V. D., and Malcolm, W. P. (2005). Pairs trading. *Quantitative Finance*, 5(3):271–276.

Fama, E. F. and French, K. R. (2023). Production of U.S. SMB and HML in the Fama-French data library. *SSRN Electronic Journal*.

Fang, L. and Peress, J. (2009). Media coverage and the cross-section of stock returns. *The Journal of Finance*, 64(5):2023–2052.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, 75(3):1327–1370.

Focardi, S. M. (2005). Clustering economic and financial time series: Exploring the existence of stable correlation conditions. *The Intertek Group*.

Fogel, F., d'Aspremont, A., and Vojnovic, M. (2014). Serialrank: Spectral ranking using seriation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Frazzini, A. and Cohen, L. (2008). Economic links and predictable returns. *Journal of Finance*, 63:1977–2011.

Gatta, F., Iorio, C., Chiaro, D., Giampaolo, F., and Cuomo, S. (2023). Statistical arbitrage in the stock markets by the means of multiple time horizons clustering. *Neural Computing and Applications*, 35:1–19.

Gervais, S., Kaniel, R., and Mingelgrin, D. H. (2001). The high-volume return premium. *The Journal of Finance*, 56(3):877–919.

Gleason, C. A. and Lee, C. M. C. (2003). Analyst forecast revisions and market price discovery. *The Accounting Review*, 78(1):193–225.

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

Gu, S., Kelly, B. T., and Xiu, D. (2019). Empirical asset pricing via machine learning. *Chicago Booth Research Paper*, (18-04):2018–09.

Gyurkó, L. G., Lyons, T., Kontkowski, M., and Field, J. (2013). Extracting information from the signature of a financial data stream.

Hause, J. C. (1971). Spectral analysis and the detection of lead-lag relations. *The American Economic Review*, 61(1):213–217.

He, Z. and Krishnamurthy, A. (2013). Intermediary asset pricing. *American Economic Review*, 103(2):732–70.

Hendershott, T., Livdan, D., and Rösch, D. (2020). Asset pricing: A tale of night and day. *Journal of Financial Economics*, 138(3):635–662.

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2011). Limited investor attention and stock market misreactions to accounting information. *The Review of Asset Pricing Studies*, 1(1):35–73.

Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184.

Hou, K. (2007). Industry Information Diffusion and the Lead-lag Effect in Stock Returns. *The Review of Financial Studies*, 20(4):1113–1138.

Huang, S., Lee, C. M., Song, Y., and Xiang, H. (2022). A frog in every pan: Information discreteness and the lead-lag returns puzzle. *Journal of Financial Economics*, 145(2, Part A):83–102.

Hubert, L. & Arabie, P. (1985). Comparing partitions. pages 193–218.

Israeli, D., Kaniel, R., and Sridharan, S. A. (2022). The real side of the high-volume return premium. *Management Science*, 68(2):1426–1449.

Kakushadze, Z. (2016). 101 formulaic alphas.

Kaniel, R., Ozoguz, A., and Starks, L. (2012). The high volume return premium: Cross country evidence. *Journal of Financial Economics*, 103(2):255–279.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154. Curran Associates, Inc.

Kelly, B., Malamud, S., and Pedersen, L. H. (2023). Principal portfolios. *The Journal of Finance*, 78(1):347–387.

Kelly, B. T., Malamud, S., and Zhou, K. (2022). The virtue of complexity in return prediction. Working Paper 30217, National Bureau of Economic Research.

Knyazev, A. (2001). Toward the optimal preconditioned eigensolver: Locally optimal block

preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 23(2):517–541.

Knyazev, A. (2017). Signed laplacian for spectral clustering revisited.

Kothari, S., So, E. C., and Verdi, R. S. (2016). Analysts' forecasts and asset pricing: A survey. *Annual Review of Financial Economics*, 8:197–219.

Kunegis, J., Schmidt, S., Lommatzsch, A., Lerner, J., De Luca, E. W., and Albayrak, S. (2010). Spectral analysis of signed graphs for clustering, prediction and visualization. *SDM*, 10.

León, D., Aragón, A., Sandoval, J., Hernández, G., Arévalo, A., and no, J. N. (2017). Clustering algorithms for risk-adjusted portfolio construction. *Procedia Computer Science*, 108:1334–1343. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

Lerman, A., Livnat, J., and Mendenhall, R. R. (2008). The high-volume return premium and post-earnings announcement drift. *SSRN Electronic Journal*.

Li, Y., Liu, C., Wang, T., and Sun, B. (2021). Dynamic patterns of daily lead-lag networks in stock markets. *Quantitative Finance*, 21(12):2055–2068.

Lo, A. W. and MacKinlay, A. C. (1990). When are contrarian profits due to stock market overreaction? *The Review of Financial Studies*, 3(2):175–205.

Lou, D., Polk, C., and Skouras, S. (2019). A tug of war: Overnight versus intraday expected returns. *Journal of Financial Economics*, 134(1):192–213.

Loudis, J. A. (2024). Stock price reactions to the information and bias in analyst-expected returns. *The Accounting Review*, 99(4):281–313.

Lyons, T. (2014). *Rough paths, signatures and the modelling of functions on streams*. Kyung Moon SA.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91.

McGowan, M. (2010). The rise of computerized high frequency trading: Use and controversy. *Duke Law and Technology Review*, 16.

Menzly, L. and Ozbas, O. (2009). Market segmentation and cross-predictability of returns. *Journal of Finance*, 65:1555–1580.

Merton, R. C. (1987). A simple model of capital market equilibrium with incomplete information. *The Journal of Finance*, 42(3):483–510.

Mikhail, S., Walther, S., and Willis, C. (1999). Does forecast accuracy matter to security analysts? *The Accounting Review*, 74:185–200.

Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Ng, A. Y., Jordan, M. I., Weiss, Y., et al. (2002). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856.

Palley, A. B., Steffen, T. D., and Zhang, X. F. (2024). The effect of dispersion on the informativeness of consensus analyst target prices. *Management Science*.

Parsons, C. A., Sabbatucci, R., and Titman, S. (2020). Geographic Lead-Lag Effects. *The Review of Financial Studies*, 33(10):4721–4770.

Pavlidis, N. G., Plagianakos, V. P., Tasoulis, D. K., and Vrahatis, M. N. (2006). Financial

forecasting through unsupervised clustering and neural networks. *Operational Research*, 6(2):103–127.

Pohl, W., Schmedders, K., and Wilms, O. (2018). Higher order effects in asset pricing models with long-run risks. *The Journal of Finance*, 73(3):1061–1111.

Sakurai, Y., Papadimitriou, S., and Faloutsos, C. (2005). Braid: Stream mining through group lag correlations. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 599–610, New York, NY, USA. Association for Computing Machinery.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415.

Savor, P. and Wilson, M. (2014). Asset pricing: A tale of two days. *Journal of Financial Economics*, 113(2):171–201.

Scherbina, A. and Schlusche, B. (2018). Follow the Leader: Using the Stock Market to Uncover Information Flows between Firms. *Review of Finance*, 24(1):189–225.

Shi, D., Calliess, J.-P., and Cucuringu, M. (2023). Multireference Alignment for Lead-Lag Detection in Multivariate Time Series and Equity Trading. *to appear in Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF 2023); SSRN:4560780*.

Smith, D. and Desormeau, W. (2006). Optimal rebalancing frequency for stock-bond portfolios. *Journal of Financial Planning*, 19:52–63.

Sortino Frank A., P. L. N. (1994). Performance measurement in a downside risk framework.

Stickel, S. E. (1992). Reputation and performance among security analysts. *The Journal of Finance*, 47(5):1811–1836.

Tamura, H. (2002). Individual-analyst characteristics and forecast error. *Financial Analysts Journal*, 58(4):28–35.

Tang, W., Xu, X., and Zhou, X. Y. (2021). Asset selection via correlation blockmodel clustering.

Tola, V., Lillo, F., Gallegati, M., and Mantegna, R. N. (2008). Cluster analysis for portfolio optimization. *Journal of Economic Dynamics and Control*, 32(1):235–258. Applications of statistical physics in economics and finance.

Tolun Tayalı, S. (2020). A novel backtesting methodology for clustering in mean-variance portfolio optimization. *Knowledge-Based Systems*, 209:106454.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Wang, K., Xu, X., and Zhou, X. Y. (2022). Variable clustering via distributionally robust nodewise regression.

Womack, K. L. (1996). Do brokerage analysts' recommendations have investment value? *Journal of Finance*.

Yan, J. and Yu, J. (2023). Cross-stock momentum and factor momentum. *Journal of Financial Economics*, 150(2):103716.

Zaharudin, K. Z., Young, M. R., and Hsu, W.-H. (2022). High-frequency trading: Definition, implications, and controversies. *Journal of Economic Surveys*, 36(1):75–107.

Zhang, T., Li, Y., Jin, Y., and Li, J. (2020). Autoalpha: an efficient hierarchical evolutionary algorithm for mining alpha factors in quantitative investment.

Zhang, Y., Cucuringu, M., Shestopaloff, A. Y., and Zohren, S. (2023). Dynamic Time Warping for Lead-Lag Relationships in Lagged Multi-Factor Models. *to appear in Proceedings of the Fourth ACM International Conference on AI in Finance (ICAIF 2023); SSRN:4572554*.

Ziegler, H., Jenny, M., Gruse, T., and Keim, D. A. (2010). Visual market sector analysis for financial time series data. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 83–90. IEEE.

A | Appendix 1

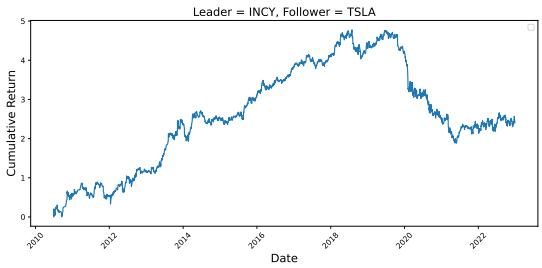
A.1 Appendix 1: Notable lead-lag pairs as motivating examples

We present several motivating examples of pairwise lead-lag relationships that we detect. Some examples show the time-varying property of lead-lag relationships while others present examples of lead-lag relationships that cannot easily reconcile with what the literature discovered in the past.

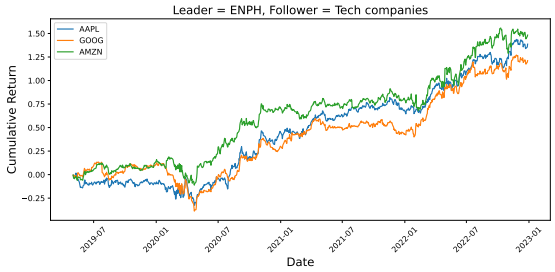
For each proposed lead-lag pair, we consider a pairs trading strategies where one uses the previous time period return of the leader as an indicator to buy or sell the follower at the current time period.

Figure A.1 presents returns of pairs trading strategies that illustrate the time-varying property of lead-lag relationships. Example A.1a shows that a ten-year lead-lag relationship between INCY, a pharmaceutical company and Tesla breaks in October 2019 when Tesla had a surprisingly good earnings announcement. Example A.1b shows that a collection of lead-lag relationships between ENPH, a Californian energy company and several technology firms including AAPL, GOOG, and AMZN started to appear in mid-2020, coinciding COVID. These two motivating examples show that lead-lag relationships can appear and disappear between pairs of stocks, which suggests that it is necessary to detect, instead of assume, lead-lag relationships over rolling windows.

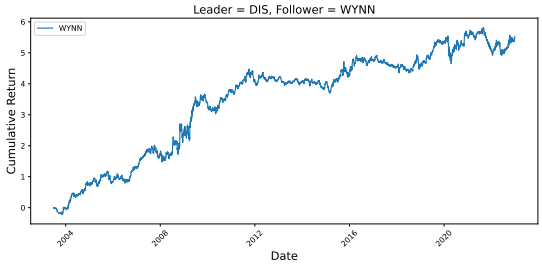
Example A.1c shows that a twenty-year lead-lag relationship between Disney and WYNN, a casino and resort company. Wynn Resorts has larger market cap than Disney and there



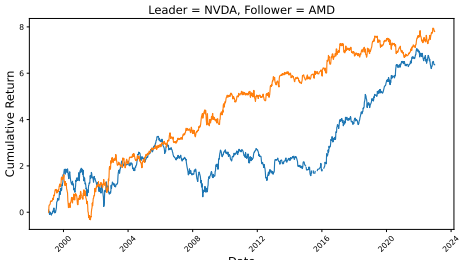
(a) Returns of INCY-TSLA lead-lag pair



(b) Returns of ENPH-Tech lead-lag pairs



(c) Returns of DIS-WYNN lead-lag pair



(d) Returns of NVDA-AMD lead-lag pair

Figure A.1: Motivating examples of pairwise lead-lag relationships.

is no clear economic links between the two companies; therefore, one cannot easily explain this lead-lag pair using lead-lag relationships discovered in the literature. On the other hand, Example A.1d shows a surprising lead-lag relationship between NVIDIA and AMD, two GPU companies. This example is built on a weekly frequency, i.e., the pairs trading strategy is executed only weekly. It is surprising that a simple weekly pairs trading strategy (orange) between two firms that are clearly economically linked is still profitable in 2023.

A.2 Appendix 2: Introduction to Signature and Lévy-area

We provide a brief introduction to the concept of Signature and Lévy-area in this section, see Lyons (2014), Gyurkó et al. (2013), and Chevyrev and Kormilitzin (2016) for more details.

First, we define the concept of a real, continuous path.

Definition 1 *A real, continuous path defined on the interval* $[0,T]$ *is a continuous function* $f:[0,T]\to\mathbb{R}$.

For a path $X_t=\left(X_t^1,\cdots,X_t^d\right)$ where each X_t^j is a real and continuous path, consider the d^k dimensional set I^0 with elements $(i_i,\cdots,i_k)$ with $k\geq 0$ and $i_m\in\{1,\cdots,d\}$ for $m=1,\cdots,k$. The signature $S_{s,t}(X)$ of the path X over a time interval $[s,t]$ is a map from X to a sequence $\left(S(X)_{s,t}^I\right)_{I\in I^0}$, where

$$S(X)_{s,t}^I=\int_{s<u_1<u_2<\cdots<u_k<t} dX_{u_1}^{i_i}\,dX_{u_2}^{i_2}\cdots dX_{u_k}^{i_k}\,. \tag{A.1}$$

Any elements i_p,i_q in the set I^0 can be different or the same, and hence if the entries of $S_{s,t}(X)$ are iterated, one obtains expressions such as $S(X)_{s,t}^{1,1,2}$ or $S(X)_{s,t}^{1,2,2}$.

For a clearer intuition, note that the signature of X can be expanded as

$$S(X)_{s,t}=(1,S(X)_{s,t}^1,\cdots,S(X)_{s,t}^d,S(X)_{s,t}^{1,1},S(X)_{s,t}^{1,2},\cdots)\,. \tag{A.2}$$

Here, for the first-order terms in $S(X)_{s,t}$, it is clear that $S(X)_{s,t}^i=X_t^i-X_s^i$, and by

definition, the higher-order terms in $S(X)_{s,t}$ can also be iteratively calculated as

$$S(X)^{i,j}_{s,t} = \int_{s<a<t} S(X)^{i}_{s,a}\, dX^{j}_{a}\,. \tag{A.3}$$

The Lévy-area between two paths X_i and X_i is

$$A_{i,j} = \frac{1}{2}\left(S(X)^{i,j}_{s,t} - S(X)^{j,i}_{s,t}\right). \tag{A.4}$$

For a more intuitive understanding, Figure A.2 shows a two-dimensional continuous path with dimensions $X^1 = \{X^1_0, \cdots, X^1_3\}$ and $X^2 = \{X^2_0, \cdots, X^2_3\}$. The Lévy-area between X^1 and X^2 is $A^+ - A^-$ where A^+ and A^- are regions bounded by the path itself and the chord connecting the start and end of the path. In this example, we set $X^1 = \{0, 0.5, 2, 2.5\}$, $X^2 = \{0.5, 2, 2.5, 3.5\}$; i.e., we artificially set X^2 to lead X^1. The Lévy-area between X^1 and X^2 is negative.

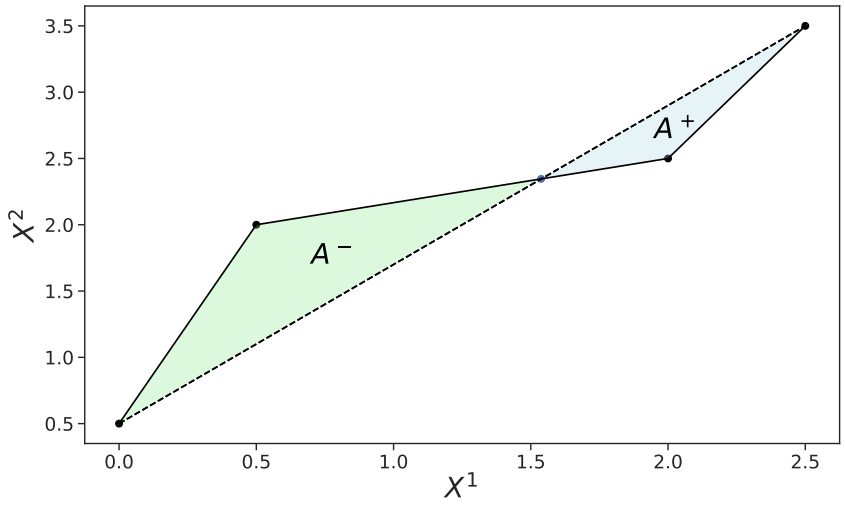


Figure A.2: Illustration of the Lévy-area between two time series X^1 and X^2

Hence, for a multi-dimensional path X, we can assess and quantify the lead-lag relationship

between two dimensions of X by calculating the signature and Lévy-area of the path along the two dimensions.

A.3 Appendix 3: Proof of Theorem 1

Proof 1 *Without loss of generality, assume $\ell > 0$. If $\ell < 0$, we re-write the regression equation by exchanging the order of X^i and X^j. Consider the Lévy-area between X_t^i and X_t^j*

$$\mathrm{A}_{i,j} = \frac{1}{2}\left(S(X)_{s,t}^{i,j} - S(X)_{s,t}^{j,i}\right) . \tag{B.1}$$

Next, write the right-hand side of (B.1) *as*

$$S(X)_{s,t}^{i,j} - S(X)_{s,t}^{j,i} = \int_{s<a<t}\int_{s<b<a} dX_b^i\, dX_a^j - \int_{s<a<t}\int_{s<b<a} dX_b^j\, dX_a^i . \tag{B.2}$$

Integrate the inner parts of the right-hand side of (B.2) *to obtain*

$$\int_{s<a<t}\int_{s<b<a} dX_b^i\, dX_a^j - \int_{s<a<t}\int_{s<b<a} dX_b^j\, dX_a^i = \int_{s<a<t}(X_a^i - X_s^i)\, dX_a^j - \int_{s<a<t}(X_a^j - X_s^i)\, dX_a^i . \tag{B.3}$$

Without loss of generality, assume that $X_s^i = X_s^j = 0$ (i.e., the return of assets at the start of time is 0), and hence we have

$$S(X)_{s,t}^{i,j} - S(X)_{s,t}^{j,i} = \int_{s<a<t} X_a^i\, dX_a^j - \int_{s<a<t} X_a^j\, dX_a^i . \tag{B.4}$$

The integration can be transformed into finite summations

$$\begin{aligned} S(X)_{s,t}^{i,j} - S(X)_{s,t}^{j,i} &= \sum_{s<a<t} X_a^i \Delta X_a^j - \sum_{s<a<t} X_a^j\, \Delta X_a^i \\ &= \sum_{s<a<t} X_a^i\,(X_a^j - X_{a-1}^j) - \sum_{s<a<t} X_a^j\,(X_a^i - X_{a-1}^i) \\ &= \sum_{s<a<t} -X_a^i\, X_{a-1}^j - \sum_{s<a<t} -X_a^j\, X_{a-1}^i \end{aligned} \tag{B.5}$$

because the processes X_t^i *and* X_t^j *are discrete.*

Recall that in model (4.6)*,* $X_a^i = \beta_A\, f_a + \epsilon_{1,a}$ *and* $X_{a-1}^j = \beta_B\, f_{a-1} + \tilde{\beta_B} g(f_{a-\ell}) + \epsilon_{2,a-1}$*. Substitute the above into equation* (B.5) *to obtain*

$$\sum_{s<a<t} -X_a^i X_{a-1}^j - \sum_{s<a<t} -X_a^j X_{a-1}^i = \sum_{s<a<t} (\beta_A\, f_a + \epsilon_{1,a})\,(\beta_B\, f_{a-1} + \tilde{\beta_B} g(f_{a-\ell}) + \epsilon_{2,a-1}) - \sum_{s<a<t} (\beta_B\, f_a + \tilde{\beta_B} g(f_{a-\ell+1}) + \epsilon_{2,a})\,(\beta_A\, f_{a-1} + \epsilon_{1,a-1})\,. \tag{B.6}$$

Next, take expectation on both sides to reduce equation (B.6) *to*

$$\begin{aligned}\mathbb{E}\left[\sum_{s<a<t} X_a^j\, X_{a-1}^i - \sum_{s<a<t} X_a^i\, X_{a-1}^j\right] &= \mathbb{E}\left[\sum_{s<a<t} (\beta_A\, \beta_B\, f_a\, f_{a-1} + \beta_A\, \tilde{\beta_B}\, f_a\, g(f_{a-\ell}))\right] \\ &\quad - \mathbb{E}\left[\sum_{s<a<t} (\beta_A\, \beta_B\, f_a\, f_{a-1} + \beta_A\, \tilde{\beta_B}\, f_a\, g(f_{a-\ell+1}))\right] \\ &= \beta_A\, \tilde{\beta_B}\, \mathbb{E} \sum_{s<a<t} (f_a\, g(f_{a-\ell}) - f_a\, g(f_{a-\ell+1})).\end{aligned} \tag{B.7}$$

Note that we assumed the data generating process of f_a *to be i.i.d. with mean 0, so the expected value of* $\beta_A\, \beta_B\, f_a\, f_{a-1}$ *and* $f_a\, g(f_{a-\ell})$ *are just 0. Additionally, if* $\ell \neq 1$*, the expected value of* $f_a\, g(f_{a-\ell+1})$ *will also be 0.*

Therefore, equation (B.7) *becomes just* $\beta_A\, \tilde{\beta_B} C \mathbb{1}(\ell = 1)$ *where* $C = \mathbb{E}(\sum_{s<a<t} f_a\, g(f_a))$*. Because we assume* $\mathbb{E}(f_a\, g(f_a)) \geq 0$ *for the given function* g*, at this point, we know that the Lévy-area between time series* i *and* j *is non-negative and proportional to* $\beta_A\, \tilde{\beta_B}$*, which proves the theorem.*

A.4 Appendix 4: Alternative Ranking Methods

Throughout the paper, we identified leaders and followers based on a column average ranking on the lead-lag matrix. Here, we use SpringRank by De Bacco et al. (2018), Serial Ranking by Fogel et al. (2014), and SyncRank by Cucuringu (2015) to construct the same portfolios as in the previous sections of the paper.

In synthetic data simulations where we test the ability of these ranking algorithms to identify lead-lag relationships with various levels of noise, we observe that all three alternative ranking methods above can detect lead-lag relationships. In particular, SpringRank produces very similar rankings as those by the method we used in the main parts of this paper (i.e., ranking by column average); on the other hand, Serial Ranking and SyncRank are more sensitive to the level of noise, the size of the lag than SpringRank, and ranking by column average.

The observations in our synthetic data simulation suggest that ranking by column average is the best choice because of its interpretability and performance. Below we report the performance of alternative ranking methods on portfolios built with the top quantile of stocks in market capitalization on each trading day, which is the same data as that used to construct the portfolios reported in Panel A of Table 4.1.

Table A.1 presents results for the same experiment as presented in Panel A of Table 4.1 with alternative specification of ranking methods. Panels A, B, C of Table A.1 show the results for data-driven portfolios when leaders and followers are identified with SpringRank, Serial Ranking, and SyncRank, respectively.

Table A.1 shows that lead-lag relationships identified by the data-driven methods presented

Table A.1: Performances of Alternative Ranking Methods

Panel A: SpringRank					
	Ann. Return (%)	Return (bps/day)	Vol. (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	19.7	7.13	0.48	2.37	16.9
Avg Cross-Cor	27.9	9.76	0.71	2.21	28.6
Lévy-area	24.9	8.82	0.59	2.38	24.7
Panel B: Serial Ranking					
	Ann. Return (%)	Return (bps/day)	Vol. (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	14.4	5.33	0.47	1.78	32.5
Avg Cross-Cor	13.5	5.04	0.48	1.68	27.5
Lévy-area	14.2	5.27	0.48	1.76	31.4
Panel C: SyncRank					
	Ann. Return (%)	Return (bps/day)	Vol. (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	12.7	4.77	0.49	1.57	28.33
Avg Cross-Cor	12.7	4.76	0.47	1.58	29.37
Lévy-area	11.5	4.33	0.48	1.41	30.48

in this paper are robust under various ranking methods. While there is a difference between the results of Table A.1 and Table 4.1, the portfolios still remain economically significant.

Portfolio performances using SpringRank are similar to the results in Table 4.1, while the results for Serial Ranking and SyncRank are not as good as those reported in Table 4.1 where the assets are ranked by column average. This is consistent with the results in the synthetic data simulations where SyncRank and Serial ranking are deemed less efficient than SpringRank and column average ranking in identifying lead-lag relationships.

A.5 Appendix 5: Summary Statistics

Table A.2: Average number of Firms Traded

Average number of firms traded						
	1963-1969	1970-1979	1980-1989	1990-1999	2000-2009	2010-2023
nondurables	7.1	28.61	38.48	44.62	37.36	33.68
durables	10.5	16.7	17.5	20.11	14.8	11.94
manufacturing	38.26	70.21	75.36	81.49	71.65	49.63
energy	14.78	34.5	37.11	29.79	40.9	41.01
chemicals	11.52	20.85	25.98	22.69	17.99	21.36
business equipment	19.61	26.54	57.13	111.04	161.54	94.82
telecommunications	3.17	6.09	13.12	26.11	35.86	21.25
utilities	5.19	44.71	63.2	44.87	34.28	33.56
shops	4.99	19.87	39.45	70.37	77.5	67.2
healthcare	7.37	20.53	32.59	49.69	66.41	47.89
finance	2.99	33.97	66.67	96.6	105.23	74.45
others	17.62	34.03	48.06	53.11	74.91	94.1
all	142.9	356.59	514.64	650.49	738.42	592.56

Table A.3: Mean (Volatility) of Price (Dollars) of Firms Traded

Mean (volatility) of price (dollars) of firms traded						
	1963–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2023
nondurables	49.23 (22.79)	34.72 (21.47)	41.82 (19.96)	43.85 (22.93)	43.08 (35.27)	72.72 (71.86)
durables	65.44 (30.73)	33.08 (21.29)	40.46 (22.58)	43.55 (18.80)	42.66 (24.11)	65.52 (43.24)
manufacturing	61.08 (43.58)	35.42 (23.72)	39.27 (19.28)	44.14 (20.37)	48.50 (26.03)	86.13 (82.22)
energy	66.53 (101.30)	43.07 (35.81)	36.96 (21.34)	43.28 (24.08)	46.10 (21.92)	50.96 (35.49)
chemicals	73.55 (47.34)	46.93 (31.87)	40.32 (23.51)	50.90 (23.01)	48.27 (18.96)	93.90 (70.89)
business equipment	87.65 (98.26)	49.43 (57.96)	39.23 (37.15)	36.46 (22.16)	36.14 (37.49)	126.63 (246.15)
telecommunications	55.80 (24.14)	34.59 (18.91)	54.01 (55.79)	52.26 (72.23)	31.67 (21.97)	67.37 (107.26)
utilities	39.91 (18.53)	22.67 (9.56)	24.94 (10.72)	31.03 (10.77)	38.38 (19.47)	51.87 (30.92)
shops	49.20 (22.66)	32.60 (22.91)	34.17 (17.15)	34.66 (16.87)	36.83 (18.97)	107.97 (160.46)
healthcare	54.83 (20.30)	43.99 (27.57)	41.61 (24.81)	42.00 (23.64)	46.45 (25.92)	118.17 (113.66)
finance	54.03 (29.95)	29.97 (16.26)	37.32 (17.97)	152.43 (2380.89)	874.92 (8546.08)	3352.74 (30629.77)
others	57.35 (29.39)	29.82 (21.91)	32.53 (19.13)	38.36 (23.12)	47.58 (57.12)	122.28 (307.53)
all	63.83 (59.15)	35.38 (28.69)	36.87 (24.07)	56.83 (918.69)	160.24 (3239.38)	507.99 (10912.13)

Table A.4: Mean (Volatility) of Daily Stock Returns (Percent)

Mean (volatility) of daily stock returns (percent)						
	1963-1969	1970-1979	1980-1989	1990-1999	2000-2009	2010-2023
nondurables	0.05 (1.64)	0.03 (2.1)	0.1 (2.11)	0.05 (1.87)	0.02 (2.12)	0.06 (1.92)
durables	0.04 (2.08)	0.0 (2.51)	0.08 (2.35)	0.08 (2.13)	-0.0 (2.97)	0.05 (2.28)
manufacturing	0.06 (1.83)	0.03 (2.25)	0.06 (2.28)	0.06 (2.24)	0.01 (3.05)	0.05 (2.26)
energy	0.07 (1.46)	0.05 (2.35)	0.05 (2.75)	0.02 (2.33)	0.04 (3.27)	0.03 (2.97)
chemicals	0.04 (1.46)	0.04 (2.02)	0.07 (2.15)	0.07 (1.73)	0.04 (2.46)	0.06 (1.95)
business equipment	0.1 (2.51)	0.02 (2.72)	0.04 (2.94)	0.1 (3.75)	-0.04 (4.87)	0.06 (2.5)
telecommunications	0.04 (1.77)	0.04 (2.2)	0.09 (2.14)	0.09 (2.75)	-0.07 (4.09)	0.05 (2.15)
utilities	0.05 (1.04)	0.04 (1.56)	0.07 (1.62)	0.04 (1.3)	0.03 (2.68)	0.05 (1.71)
shops	0.05 (1.7)	0.01 (2.61)	0.08 (2.47)	0.07 (2.67)	0.02 (2.8)	0.06 (2.36)
healthcare	0.06 (1.62)	0.02 (2.21)	0.07 (2.44)	0.07 (2.99)	0.03 (3.21)	0.06 (2.29)
finance	0.11 (2.08)	0.02 (2.3)	0.06 (2.16)	0.09 (2.2)	0.03 (3.01)	0.06 (2.18)
others	0.11 (2.37)	0.04 (2.82)	0.06 (2.64)	0.04 (2.75)	0.01 (3.61)	0.05 (2.93)
all	0.07 (1.93)	0.03 (2.3)	0.07 (2.36)	0.07 (2.64)	0.01 (3.57)	0.06 (2.42)

Table A.5: Mean (Volatility) of Volume (Thousand Shares) Traded

Mean (volatility) of volume (thousand shares) traded						
	1963–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2023
nondurables	15.89 (25.75)	24.53 (31.66)	168.69 (330.68)	504.54 (903.89)	2023.79 (2927.21)	3574.74 (4489.26)
durables	30.76 (34.79)	40.15 (51.52)	263.06 (427.55)	766.86 (1207.08)	4650.96 (11069.88)	7021.39 (17920.51)
manufacturing	16.83 (19.65)	29.08 (38.63)	154.66 (263.95)	423.70 (743.63)	2348.53 (5550.60)	4474.21 (10580.17)
energy	17.61 (25.34)	39.63 (50.18)	232.54 (421.07)	579.24 (655.04)	2984.50 (4875.58)	5660.98 (7504.94)
chemicals	12.09 (14.81)	31.23 (36.93)	175.51 (314.65)	430.95 (590.71)	2127.48 (2812.24)	2876.99 (3522.38)
business equipment	27.47 (33.47)	38.93 (42.24)	244.00 (372.40)	1142.64 (2231.66)	5988.36 (12347.37)	6662.66 (13189.96)
telecommunications	32.78 (40.32)	46.20 (61.41)	349.05 (813.14)	1096.04 (1829.91)	6417.22 (13353.32)	9018.87 (15634.01)
utilities	10.33 (9.76)	24.93 (33.29)	182.96 (886.38)	317.95 (411.83)	1961.25 (2639.03)	3572.77 (4873.39)
shops	17.81 (25.55)	33.44 (45.71)	211.61 (334.01)	582.80 (872.33)	2528.56 (3470.74)	3402.12 (4607.55)
healthcare	11.89 (14.47)	32.18 (39.65)	207.85 (278.74)	662.55 (900.29)	2944.18 (5148.06)	3691.78 (6760.03)
finance	17.18 (30.36)	29.42 (40.17)	182.96 (290.99)	457.31 (704.05)	3335.61 (11066.80)	6541.94 (25201.02)
others	21.22 (26.24)	33.91 (50.85)	163.86 (234.20)	474.90 (826.97)	2307.92 (3657.88)	4404.77 (8344.51)
all	19.42 (26.00)	31.80 (42.36)	196.67 (449.73)	634.27 (1244.02)	3592.95 (8640.04)	5019.04 (12594.18)

Table A.6: Mean (Volatility) of Daily Stock Turnovers

Mean (volatility) of daily stock turnovers						
	1963–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2023
nondurables	1.33 (3.02)	1.06 (1.59)	2.98 (4.93)	3.12 (5.28)	6.90 (9.99)	9.91 (13.75)
durables	2.68 (6.04)	1.44 (2.45)	3.55 (4.80)	3.82 (4.84)	10.31 (11.70)	12.32 (11.49)
manufacturing	2.03 (5.31)	1.65 (9.80)	3.34 (4.54)	5.20 (9.20)	13.80 (30.27)	13.20 (17.54)
energy	0.91 (2.39)	1.60 (2.71)	2.91 (3.81)	4.82 (7.11)	13.86 (13.07)	18.15 (19.28)
chemicals	0.97 (2.30)	1.17 (1.53)	3.16 (5.08)	3.12 (3.32)	7.43 (9.44)	9.11 (9.35)
business equipment	4.97 (8.81)	2.80 (5.05)	6.26 (8.09)	16.43 (22.37)	20.50 (25.05)	14.23 (21.51)
telecommunications	1.06 (1.60)	1.90 (4.49)	3.96 (6.78)	6.16 (12.19)	11.34 (16.35)	11.65 (13.62)
utilities	0.36 (0.33)	0.79 (0.98)	2.40 (7.41)	2.21 (2.93)	6.72 (7.56)	7.72 (6.64)
shops	1.55 (3.30)	1.90 (3.57)	4.29 (7.39)	6.75 (12.51)	12.71 (14.89)	15.24 (22.20)
healthcare	0.75 (1.06)	1.23 (2.64)	4.19 (6.22)	9.15 (15.82)	13.28 (20.70)	10.11 (17.85)
finance	2.95 (5.64)	1.36 (2.99)	3.41 (4.91)	4.03 (6.86)	8.52 (15.54)	9.28 (14.34)
others	3.78 (5.36)	2.30 (3.93)	4.32 (6.06)	6.97 (13.98)	16.50 (21.51)	18.04 (29.81)
all	2.32 (5.39)	1.57 (5.13)	3.73 (6.12)	7.11 (13.70)	13.61 (20.62)	13.24 (20.30)

A.6 Appendix 6: Alternative Parameters

Table A.7: Performances of Lead-lag Portfolios - Alternative Hyperparameters

	Compound Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Panel A: Global Lead-Lag Portfolios - 40% Leaders and Followers					
Max Cross-Cor	18.36	6.70	0.44	2.43	18.9
Avg Cross-Cor	21.37	7.69	0.48	2.54	18.2
Lévy-area	24.26	8.62	0.51	2.70	17.8
Market Cap	0.08	0.03	0.46	0.01	91.75
Turnover	8.90	3.38	0.35	1.52	13.93
Panel B: Clustered Lead-lag Portfolios - Hermitian Clustering					
	Compound Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	15.73	5.80	0.38	2.38	12.46
Avg Cross-Cor	21.74	7.81	1.18	1.05	38.20
Lévy-area	24.31	8.64	0.39	3.43	9.19
Panel C: Clustered Lead-lag Portfolios - 40% Leader and Lagger					
	Compound Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	15.06	5.57	0.35	2.42	11.75
Avg Cross-Cor	17.23	6.31	0.40	2.52	20.39
Lévy-area	19.65	7.12	0.38	2.99	11.51
Industry	6.64	2.55	0.38	1.06	26.9
Panel D: Clustered Lead-lag Portfolios - 40% Leader and Lagger, Hermitian Clustering					
	Compound Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	16.08	5.92	0.37	2.51	14.77
Avg Cross-Cor	20.98	7.56	1.14	1.05	39.49
Lévy-area	22.70	8.12	0.37	3.45	10.07
Panel E: Global Lead-lag Portfolios - 20% Leader and Lagger, 30 Day Look-back Window					
	Compound Return (%)	Return (bps/day)	Volatility (%)	Sharpe Ratio	Max Drawdown (%)
Max Cross-Cor	29.60	10.29	0.74	2.21	27.48
Avg Cross-Cor	22.72	8.13	0.51	2.19	16.91
Lévy-area	23.64	8.42	0.59	2.61	16.72

A.7 Appendix 7: Composition of Lead-lag Portfolios

Table A.8: Compositions of Lead-Lag Portfolios

Panel A: Characteristics for Leaders and Followers in the MaxCor Portfolio

	Leader Avg Percentile (%)	Follower Avg Percentile (%)	Permutation Test P-value
Market Cap	57.48	58.66	0.041
Volume	53.95	53.77	0.823
Turnover	39.09	37.95	0.145
Return	50.16	50.19	0.588
Price	57.22	58.87	0.045
Shares Outstanding	63.16	65.28	0.013

Panel B: Characteristics for Leaders and Followers in the AvgCor Portfolio

	Leader Avg Percentile (%)	Follower Avg Percentile (%)	Permutation Test P-value
Market Cap	57.99	59.26	0.019
Volume	53.67	53.61	0.936
Turnover	38.77	36.53	0.003
Return	50.17	50.19	0.759
Price	57.74	59.61	0.019
Shares Outstanding	63.28	66.35	<0.0001

A.8 List of Features

Table A.9: Constructed Features and Parameters

Feature	Description	Parameters / Formula
day_change	Absolute change in price from open to close	$\frac{\text{price\_close} - \text{price\_open}}{\text{price\_open}}$
spread	Intraday price spread (proxy for volatility)	$\frac{\text{price\_high} - \text{price\_low}}{\text{price\_low}}$
return_momentum	Rolling average of returns over different time windows	Window size: 5, 3, 10, 15, 20, 30, 50
volatility	Standard deviation of daily returns over different time windows	Window size: 2, 5, 7, 10, 15, 25, 30, 50
lagged_return	Lagged returns over different time lags	Lag: 1, 2, 3, 4, 5
lagged_volume_shock	Lagged values of volume shocks over different time lags	Lag: 1, 2, 3, 4, 5
volume_shock_rolling_mean	Rolling mean of volume shocks over different time windows	Window size: 5, 10, 2, 3
volume_shock_rolling_std	Rolling standard deviation of volume shocks over different time windows	Window size: 5, 10, 2, 3
volume_shock_diff	Difference of volume shocks over different time lags	Lag: 1, 2
volume_shock_pct_change	Percentage change of volume_shock over different time lags	Lag: 1, 2
close_to_open_ratio	Ratio of closing price to opening price	$\frac{\text{price\_close}}{\text{price\_open}}$
high_to_low_ratio	Ratio of daily high price to daily low price	$\frac{\text{price\_high}}{\text{price\_low}}$
day_of_week	Day-of-week dummy variables	Dummy variable for each day (Monday to Friday)
month	Month-of-year dummy variables	Dummy variable for each month (January to December)
ADV	Average daily volume over different time windows	Window size: 5, 2, 3, 10, 15, 20, 25, 30, 50
E_ADV	Exponential average of daily volume over different time windows	Halflife: 2, 3, 5, 10, 20, 30, 60
macd	Moving Average Convergence Divergence (MACD) for various features	Short window: 12, Long window: 26, Signal window: 9
volume_roc	Rate of change of volume	$\frac{\text{volume\_pct\_change} \times 100}{1}$
intraday_vol	A proxy of intraday volatility using close, open, high, and low price	$\frac{\text{price\_close} - \text{price\_open}}{\text{price\_high} - \text{price\_low} + 0.001}$

A.9 Gradient Boosting Decision Trees and LightGBM

A.9.1 Gradient Boosting Decision Trees (GBDT)

Decision Tree is a machine learning model used for both classification and regression tasks. It recursively partitions the vector space of input features into regions and assigns a predictive model to each partition. The model is constructed through a process that optimizes certain loss functions, such as maximizing information gain or minimizing impurity.

Let X be the input feature space, and Y be the corresponding space of predictive targets. A decision tree is a mapping $f : X \to Y$ through a series of binary decisions which we call node decisions. At each node j, a decision is made based on a feature x_i and a threshold θ_j:

$$\text{Node Decision: } x_i \leq \theta_j \, .$$

This decision determines the assignment of an entry of data to the left or right child node. Therefore, every entry of input is assigned to a leaf node of the tree, i.e., a node that does not have child nodes. Next, the model assigns predicted values or class labels to each leaf node based on the task of the model.

The construction of a decision tree involves recursively selecting the optimal features and thresholds to optimize a certain criterion or loss function. For a regression task, the criterion for each node decision is often the reduction in mean squared error (MSE)

$$\text{MSE} = \frac{1}{N_m} \sum_{i \in D_m} (y_i - \overline{y}_m)^2 .$$

Here N_m is the number of data points in node m, D_m is the set of indices of data points in node m, y_i is the target value for data point i, and $\overline{y}_m$ is the mean target value in node m.

The algorithm selects the feature x_i and threshold θ that maximizes the reduction in MSE

$$\text{Reduction in MSE} = \text{MSE}_{\text{parent}} - \left(\frac{N_{\text{left}}}{N_{\text{parent}}} \text{MSE}_{\text{left}} + \frac{N_{\text{right}}}{N_{\text{parent}}} \text{MSE}_{\text{right}} \right) .$$

Here, N_{left}, N_{right} are the number of data points in the left and right child nodes, and MSE_{left}, $\text{MSE}_{\text{right}}$ are the MSE values for the left and right child nodes.

The algorithm continues recursively until a stopping criterion is met, such as the tree has reached a maximum depth or has too few data points in some nodes. At that time, the nodes that do not have child nodes are denoted leaf nodes, and for each leaf node, we assign the predictive value to be the average of target values of data points that belong to the node.

Decision Trees provide a transparent and interpretable model, with each split representing a decision based on a specific feature and threshold. The resulting structure forms a hierarchical representation of the input space, enabling intuitive understanding of the model's decision-making process.

Gradient Boosting Decision Trees (GBDT) is an optimization process seeking the best approximation of the target variable y through a combination of decision trees. Mathematically, this is represented by the additive model

$$F(x) = \sum_{m=1}^{M} f_m(x).$$

Here, f_m denotes the m-th decision tree, and M is the total number of trees in the ensemble.

The optimization objective is to minimize a loss function, often the mean squared error (MSE) for regression tasks:

$$L(y, F(x)) = \frac{1}{2} \sum_{i=1}^{N} (y_i - F(x_i))^2 ,$$

where N is the number of observations in the dataset.

At every step of optimization, the model aims to find the optimal tree $h(x)$ that minimizes the loss:

$$h_m(x) = \arg\min_h \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + h(x_i)) .$$

In the process of fitting the model, the model updates its values on the direction of the negative gradient of the loss function, denoted as $g_m = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}$. At each iteration, each additional tree is fitted to the residual errors of the current model

$$h_m(x) = \arg\min_h \sum_{i=1}^{N} \left[y_i - F_{m-1}(x_i) - h(x_i) \right]^2 .$$

A.9.2 LightGBM

LightGBM is a recent implementation of Gradient Boosting Decision Trees. This model introduces innovative techniques for efficient and scalable gradient boosting and optimization. Its key features include histogram-based learning and the Gradient-based One-Side

Sampling (GOSS) technique. Histogram-based learning allows the model to discretize continuous features into bins and to construct histograms to speed up the training process. GOSS selectively samples instances based on their gradient values, prioritizing those with larger gradients. Both approaches accelerate convergence of the model, and hence the speed of training.

LightGBM optimizes the objective function, combining the loss function of GBDT and regularization terms

$$\text{Obj}(\theta) = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{k=1}^{K} \Omega(f_k)\,.$$

The algorithm employs a histogram-based approach for tree construction. Let B be the number of bins, and $H_{j,k}$ be the histogram for the j-th feature and k-th bin. The objective is to find the best split

$$\text{Split: } \sum_{j=1}^{J} \sum_{k=1}^{K} \left[\frac{G_{j,k}^2}{H_{j,k} + \lambda} + \frac{(G_{\text{total}} - G_{j,k})^2}{H_{\text{total}} - H_{j,k} + \lambda} - \frac{G_{\text{total}}^2}{H_{\text{total}} + \lambda} \right].$$

Here, $G_{j,k}$ and $H_{j,k}$ are the sum of gradients (g) and Hessians (h) for the j-th feature and k-th bin. The term λ is the regularization term.

LightGBM employs a leaf-wise growth strategy. The optimal leaf value $w_{j,k}$ is determined by

$$w_{j,k} = -\frac{\sum_{i \in I_{j,k}} g_i}{\sum_{i \in I_{j,k}} h_i + \lambda}\,,$$

where $I_{j,k}$ is the set of instances falling into the k-th bin of the j-th feature.

Just like GBDT, LightGBM iteratively fits decision trees to the negative gradient of the

loss function with respect to the current model's predictions. This process is guided by a learning rate (λ) that controls the contribution of each tree.

A.10 TabNet - A Deep Learning Architecture for Tabular Data

TabNet is a deep learning architecture specifically designed for handling tabular data, which is especially common in the field of financial and asset pricing studies. Unlike traditional machine learning models that may struggle to capture complex relationships within tabular data, TabNet employs a novel architecture that combines feature selection with attention mechanisms.[1] This architecture provides both predictive accuracy and interpretability compared to common benchmarks (Arik and Pfister (2021)).

Let X represent the input feature space of tabular data, and Y denote the corresponding space of predictive targets. The TabNet architecture consists of three key components: feature selection masks, attentive transformer units, and sequential decision steps.

At each decision step, TabNet learns a binary mask M over the input features, indicating which features to select. This feature selection mask is analogous to selecting features in each iteration of a GBDT. In each iteration, the selected features are assigned attention scores, which are used to update the representations through Attentive Transformer Units. For a more detailed description of the attention mechanism, see Vaswani et al. (2017). As the last step, TabNet is constructed as a sequence of decision steps. At each iteration, the model decides which features to attend to based on the learned masks, updating the representations and weights in the model.

[1]The attention mechanism, first introduced in Vaswani et al. (2017), also led to the invention of ChatGPT.

Mathematically, TabNet constructs an iterative representation $H^{(t)}$ of the features through the binary masks, attentive units, and sequential updates of the model. Let $M^{(t)}$ represent the binary mask at decision step t. The updated feature representations $H^{(t)}$ are computed as follows:

$$H^{(t)} = H^{(t-1)} + M^{(t)} \cdot \text{ATU}(H^{(t-1)}, M^{(t)})$$

where ATU denotes the Attentive Transformer Unit.

For regression tasks, apply a regression head to the last updated feature representation $H^{(T)}$ to obtain the final output O

$$O = \text{RegressionHead}(H^{(T)})$$

where RegressionHead is a regression-specific layer.

TabNet is particularly suitable for working with tabular data. FOr example, in financial studies, its feature selection masks and attention scores provide interpretability to the model by showing the importance of each feature in the prediction task. It allows us to understand which features impact the predictions the most in a similar way as the more traditional models such as linear regression.