The Battle of the Models: Modern Takes on Traditional and Machine Learning Techniques in Empirical Finance

Clint Howard

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Finance Discipline Group, UTS Business School
University of Technology Sydney

Principal supervisor: Associate Professor Vitali Alexeev
Co-supervisor: Professor Tālis J. Putniņš

October 2023

Certificate of Original Authorship

I, Clint Howard, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Finance Discipline Group at UTS Business School at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: October 11, 2023

Acknowledgments

I would like to thank my supervisors, Vitali and Talis. I am extremely grateful for the guidance and research mentorship they both provided me throughout my Ph.D. Considering my slightly unorthodox pathway into the Ph.D. program at UTS, Vitali and Talis showed great patience with me as I balanced my full-time work commitments with the completion of my thesis. I am also thankful to the support and guidance of Christina Sklibosios Nikitopoulos, who was always available for technical support and ensuring that I met all the requirements for the Ph.D. program as well as accommodating my international move along the way.

During my Ph.D. journey, I had the privilege of working for two asset management teams, Macquarie Systematic Investments and Robeco Quantitative Investments. I would like to thank my research colleagues, Samir Vanza and Steve Wong, from Macquarie. Their mentorship and research guidance had significant influence on my development as a researcher. I would also like to thank Ben Leung for his encouragement and supporting me to embark on a part-time Ph.D. program whilst continuing to work full-time.

The completion of my Ph.D. has been significantly supported by many research colleagues at Robeco. I would like to specifically thank David Blitz and Harald Lohre, who provided comprehensive feedback, critiques, suggestions, and support for my research.

I thank my parents for their everlasting encouragement of my life pursuits. Finally, I thank Pearl, whose support, patience, and understanding have kept me going throughout.

Preface

Chapters 2–4 of this thesis have been simultaneously developed as working papers. These working papers have been presented at academic conferences and to my employers during my doctoral candidacy (Macquarie Systematic Investments and Robeco Quantitative Investments). The below list presents each working paper and the relevant presentations.

1. Howard, C., Putnins, T. and Alexeev, V. (2020), To lead or to lag? Measuring asynchronicity in financial time series using dynamic time warping, *Working paper*, UTS Business School.
    - 2022 Robeco Institutional Asset Management Research Seminar Series. Rotterdam, The Netherlands.
    - 2023 Lancaster University Management School Financial Econometrics Conference. Lancaster, United Kingdom.
2. Howard, C., Putnins, T. and Alexeev, V. (2021), The index effect is not dead, it has mutated, *Working paper*, UTS Business School.
    - 2022 Robeco Institutional Asset Management Research Seminar Series. Rotterdam, The Netherlands.
3. Howard, C. (2023), Less is more? Biases and overfitting in machine learning return predictions, *Working paper*, UTS Business School.
    - 2023 Robeco Institutional Asset Management Research Seminar Series. Rotterdam, The Netherlands.
    - 2023 Inquire Europe Autumn Seminar. Cologne, Germany.
    - 2023 29th Annual Meeting of the German Finance Association Meeting. Stuttgart, Germany.

Dediscit animus sero quod didicit diu [The mind unlearns with difficulty what it has long learned.]

– Seneca

Contents

List of Tables

List of Figures

Abstract

Consensus views in finance must be continuously challenged and re-evaluated. This thesis uses new techniques and modern perspectives to challenge commonly held beliefs, both new and old, in financial markets. Across three chapters, this thesis addresses the presence of asynchronicity in financial markets, the purported death of the Standard and Poor's (S&P) index effect, and the presence of biases and overfitting in machine learning models used in asset pricing.

Chapter 2 addresses the presence of asynchronicity in financial markets and challenges the long-standing beta anomaly. Financial time series are seldom perfectly synchronized, leading to misestimation in empirical models. This chapter establishes dynamic time warping (DTW) as a measure of dynamic asynchronicity and applies DTW to improve asset pricing and price discovery models. Using DTW to correct for dynamic asynchronicity when estimating a stock's beta recovers a positive relation between market risk and return, thus helping resolve the beta anomaly. Applying DTW at intraday frequencies uncovers important price leadership dynamics in global markets that are overlooked by conventional measures of price discovery.

Chapter 3 questions the purported death of the S&P index effect. Recent research on the S&P index effect, a phenomenon where stocks added to or deleted from the S&P 500 index experience abnormal price responses, argues that it has disappeared. This chapter finds that the S&P index effect has not disappeared. Stocks added into the S&P 500 from outside the broader S&P 1500 universe still experience positive abnormal price responses. However, stocks that move between the S&P 500, S&P 400, and S&P 600 no longer exhibit abnormal price responses to index change announcements. The results connect stock price reactions to announcements of changes to the three main S&P U.S. domestic equity indexes with the impact of relative passive ownership and informed trading on the informational content of these events. The findings alleviate concerns about potential price distortions in

equity markets arising from index rebalance events alongside the growth in passive investing.

Chapter 4 critically examines the application of machine learning in asset pricing and highlights the potential biases and overfitting arising from common modeling choices. The chapter explores the performance of machine learning models trained on size-specific groups of stocks. Contrary to expectations, grouping stocks by market capitalization improves the performance of machine learning return predictions compared with models trained on the full cross-section of stocks. The superior performance of size-specific models is attributable to a lack of regularization of the target stock returns in the standard machine learning return prediction framework. The findings underscore the importance of data selection and prediction target design when training machine learning models for return prediction and serve as a cautionary reminder that machine learning requires careful guidance to reduce biases and overfitting.

In summary, this thesis challenges several commonly held views in empirical finance. The findings underscore the necessity for inventive approaches and reassessing long-standing and newly emerging beliefs. The findings demonstrate the potential of broader datasets and alternative techniques, such as DTW and neural networks, in generating novel insights and more accurate models in empirical finance.

Chapter 1

Introduction

Academics and practitioners utilize commonly accepted views, paradigms, and "rules of thumb" in their everyday work. These consensus views often exist in response to the inherent challenge in modeling real-world financial markets. Empirical models are often not sophisticated enough to fully capture the complex behavior of financial markets, often leading to empirical anomalies: contradictions between real-world behavior and the behavior predicted by financial models. In financial markets, a dogmatic approach to many of the existing puzzles and anomalies has generally been adopted. Longstanding results, while not standing up to empirical evidence, persist to this day due to their acceptance as "easy to understand" and "easy to explain." A canonical example is the capital asset pricing model (CAPM), independently proposed in the 1960s by several researchers. Despite the continuous challenges and criticism of the CAPM, it remains commonplace in finance curricula and empirical practice.

The blending of natural and social sciences is at the heart of this adherence to commonly accepted paradigms. The study of financial economics aims to develop and apply mathematical models to the real-world of financial markets. On one hand, financial economics is a social science studying the behavior of financial markets, which are themselves derivatives of human behavior and design. As humans change, the assets and markets they create and interact with also evolve. On the other hand, humans like to impose longstanding "laws" and "theories" to model real-world behavior. Nevertheless, the mathematical tools available are often not sophisticated enough to fully model ever-changing human behavior. This complexity leads to the core challenge of financial economics. Humans want to perfectly model real-world observations, yet perfectly modeling financial markets is impossible. This conflict ultimately leads to longstanding, and often dogmatic, adherence to commonly held

views, whereas in reality, the evolving nature of financial markets requires constant challenging and re-evaluation of such perspectives.

This thesis examines biases and puzzles in empirical finance that arise from both traditional and modern modeling approaches. Chapters 2–4 of this thesis each explore a different topic in empirical finance research where commonly held beliefs are prevalent. Chapter 2 resolves the CAPM beta anomaly by fully accounting for time-series asynchronicity using dynamic time warping (DTW). Chapter 3 challenges the notion that the Standard and Poor's (S&P) index effect has disappeared, finding that the S&P index effect is still present for subsets of index change announcements. Finally, Chapter 4 questions the current discipline when applying machine learning in asset pricing by demonstrating how empirical anomalies can arise from seemingly innocuous arbitrary modeling decisions.

1.1 Asynchronicity between financial time series

Asynchronicity is at the core of time-series models in financial econometrics. As markets have become faster, a prevailing view is that asynchronicity has become less problematic in empirical modeling, but this is far from accurate Although asynchronicity at lower frequencies (such as daily observations) has undoubtedly reduced, as long as latency between trading venues exists, latency will exist between common assets. The persistent nature and varying manifestations of asynchronicity continue to plague time-series models and have undue influence on model inference. Therefore, it is necessary to continue to explore methods for measuring and correcting for asynchronicity in financial models.

Chapter 2 proposes using DTW to measure asynchronicity between financial time series. DTW has a distinct advantage over prevailing methods: it estimates the lead-lag between time series for every observation in the estimation window. This key advantage allows DTW to be used in novel ways to correct for asynchronicity when comparing financial time series and to explore new ways of studying existing problems. I first use a simulation framework to demonstrate that DTW is effective at capturing stylized lead–lag structures between two time series. I subsequently use DTW to align stock returns with market returns, and then measure a stock's beta to the market on these DTW-aligned time series. By fully incorporating the dynamic asynchronicity between stock returns and market returns, the DTW-estimated betas helps resolve the longstanding beta anomaly. I also use DTW to study intraday price leadership patterns between global futures contracts. Using DTW, I uncover rich intraday lead–lag dynamics between U.S. and U.K. equity index futures that are centered around significant market operation events in the underlying equity

markets. Existing price discovery models miss such dynamics, as they are unable to provide sufficient granularity in the estimation window to uncover such patterns.

Chapter 2 also demonstrates how applying new techniques can challenge long-held consensus views around empirical results. Asynchronicity effects in trading drive the manifestation of stock betas, resulting in the beta anomaly in historical data. Although previous approaches for incorporating this asynchronicity into the measurement of beta improve the base result, they do not fully account for the dynamic nature of asynchronicity. By fully accounting for the dynamic lead–lag effects, a more accurate beta estimate can be obtained. Ultimately, DTW is shown to be a suitable method for measuring and correcting for dynamic asynchronicity between financial time series.

1.2 Is the S&P index effect dead?

One of the core features of financial markets is their self-learning nature. As academics and practitioners collectively learn and disseminate research around financial markets, participants incorporate this information into their behavior when operating in these markets. Academic literature can reveal an empirical observation in historical data, but there is no guarantee that this observation will manifest in the future realizations of the data. The commonly known S&P index effect is a key example of this phenomenon. Initial research (Harris and Gurel, 1986; Shleifer, 1986; Jain, 1987; Dhillon and Johnson, 1991; Lynch and Mendenhall, 1997; Chen, Noronha and Singal, 2004) showed that stocks experience abnormal returns when added to or deleted from the S&P 500 index and that this pattern could be exploited for profit. New results (Kamal, Lawrence, McCabe and Prakash, 2012; Kim, Li and Perry, 2017; Bender, Nagori and Tank, 2019; Bennett, Stulz and Wang, 2020), using an updated sample of index announcements, find that the S&P index effect has disappeared. Stocks no longer experience statistically significant abnormal returns when added to or deleted from the S&P 500 index. However, this claim of the death of the S&P index effect has coincided with the enormous growth in passive investing and the amount of assets passively following the S&P 500 index. With such a significant growth in assets that mechanically track the S&P 500 index, the economic prior suggests that the S&P index effect should still exist, creating a puzzling observation of the death of the index effect.

Chapter 3 examines the S&P index effect by collecting a complete sample of S&P 500, S&P 400, and S&P 600 index change announcements and tracking the internal movements that occur between these indexes. By jointly considering the abnormal return responses of stocks to index change announcements for the three S&P indexes,

I show that the S&P index effect is not dead. Rather, it is the migrations between S&P indexes that no longer experience significant abnormal price responses when index changes are announced. Stocks that are added from outside the broader S&P 1500 universe to one of the three S&P indexes still experience significant abnormal return responses when such a change is announced. By measuring the changing distribution in passive ownership between large capitalization and small capitalization stocks, I show that the S&P index effect is alive and well.

Chapter 3 further demonstrates that different approaches to studying the same problem can yield different conclusions. By replicating original studies with newer data, the original results on the existence of the S&P index effect can be discarded if the market context of the new results is not acknowledged. However, by considering how changes in market structure (such as the growth of passive investing) could impact index changes, richer insights on the S&P index effect can be obtained, complementing, and extending earlier results.

1.3 Biases and overfitting in cross-sectional machine learning models

The application of machine learning models across numerous disciplines has seen significant success in recent years. Seminal papers applying machine learning to asset pricing demonstrate the strength and superiority of machine learning models when using large sets of cross-sectional asset pricing characteristics to predict future excess returns across various asset classes. However, with such rapid growth in the literature, and the applied approach of trial-and-error for estimating these models, a rigorous understanding of how these models operate in the asset pricing domain has been understudied.

Chapter 4 critically examines the current application of machine learning models in the asset pricing literature. By imposing an economic prior on the relationship between market capitalization and future excess returns, economically significant improvements over existing approaches are obtained. By training group-specific machine learning models to predict stock returns, these model predictions outperform those trained on the entire cross-section of stocks. This result is counter-intuitive to the commonly held belief that "the more data, the better" for machine learning models in return prediction. I show how this performance improvement should not be fully attributed to the imposed economic prior around group-specific asset pricing characteristics. Instead, the gain arises predominantly from a lack of regularization in the standard machine learning model design for

predicting stock returns. This lack of regularization induces overfitting toward predicting returns for small stocks in cross-sectional machine learning models. By recognizing and correcting for this lack of regularization, similar performance gains as group-specific machine learning models can be achieved without the added computational complexity of training separate models.

Chapter 4 also demonstrates that even for more modern techniques, commonly held views and practices around these techniques should be challenged and continuously evaluated. It is detrimental to adhere to widely set empirical methods, particularly in machine learning, without questioning the economic rationale backing each decision inherent to the method. The high dimensionality of modeling decisions in machine learning means that a cautious and guided approach to investigating and comparing results from the use of machine learning in asset pricing is fundamental to its ongoing success.

1.4 Thesis outline

This thesis consists of three distinct studies in empirical finance:

i. A new measure of asynchronicity in financial time series and two empirical applications for beta estimation and price discovery (Chapter 2)

ii. The changing nature of the S&P index effect (Chapter 3)

iii. Biases and overfitting when training machine learning models for empirical asset pricing (Chapter 4)

In Chapter 5, the findings are summarized, and future research directions are presented.

Chapter 2

To lead or to lag? Measuring asynchronicity in financial time series using dynamic time warping

2.1 Introduction

Asynchronicity in financial time series poses a significant challenge in the study of financial markets. Lo and MacKinlay (1990) demonstrate that market frictions can impede information transmission, leading to lead–lag patterns between observable asset time-series data. Frictions owing to infrequent trading (Cohen, Hawawini, Maier, Schwartz and Whitcomb, 1983), information flowing from large stocks to smaller stocks (DeMiguel, Nogales and Uppal, 2014), and even the physical distance between trading venues (Laughlin, Aguirre and Grundfest, 2014) can all contribute to asynchronicity between observed time-series data. This asynchronicity between time series can cause errors in inference in financial models. For example, it can result in price discovery models missing important dynamics within the estimation window and the misestimation of asset covariance owing to asynchronicity in the underlying time series. Empirical models commonly assume that observations of multiple time series occur contemporaneously or with a fixed time lag. However, the dynamic nature of asynchronicity poses a significant challenge that extant econometric frameworks struggle to account for. In this chapter, I use DTW to measure and correct for dynamic asynchronicity between financial time series in asset pricing and price discovery models.

DTW is an alignment algorithm, first utilized in speech recognition (Sakoe and Chiba, 1978; Keogh and Pazzani, 2002), that measures the similarity between two

time series that may vary in speed. The strength of DTW is the ability to flexibly capture leading and lagging patterns between two or more time series. This feature of DTW is well-suited to measuring asynchronicity between financial time series and accounting for the time-varying nature of the asynchronicity. The use of DTW does not require structural assumptions of the expected behavior of the lead–lag. Instead, the algorithm itself freely uncovers any dynamic lead–lag structures that are present between the two time series. As the origins of DTW in signal processing were due to signal distortion which arose due to measurement differences and media in which the signals travel, there is a question around the validity of using DTW in financial markets settings. Ultimately, lead–lag in financial markets can arise from many sources, such as differences in information processing or latency between trading venues. Such differences are analogous to the signal processing and suggest that lead–lag in financial markets are likely suited to be captured by DTW.

First, I validate DTW's ability to recover lead–lag structures across several simulated time-series scenarios. The noise, volatility, and lead–lag structures between time series are explicitly controlled in four simulated scenarios. Across these simulated lead–lag scenarios, DTW successfully recovers the induced lead–lag patterns with an average mean absolute error (MAE) of 5.9%. The success of DTW in recovering the lead–lag pattern is primarily a function of the noise contaminating the time series, and the estimation error is stable for varying levels of the fundamental volatility of the simulated time series. The stability of estimation error in DTW at different fundamental volatility levels is an important result, as it suggests that DTW can be used for inherently volatile assets. To assess the practical relevance of DTW, two empirical applications in asset pricing and price discovery are investigated.

Real-world asynchronicity in financial time series occurs across different frequencies, from macro to the micro. I explore two empirical applications to demonstrate DTW's relevance for financial time series, one at the daily frequency and one at the intraday frequency. The first application uses DTW to measure a stock's beta to the market. Under the CAPM, a stock's beta is estimated using contemporaneous stock and market returns. The CAPM predicts a positive relation between a stock's beta and expected returns. Several studies (Reinganum, 1981; Lakonishok and Shapiro, 1986; Fama and French, 1992) present empirical evidence that there is, in fact, a negative relation between risk and return, contradicting the predictions of the CAPM and giving rise to the beta anomaly. The literature has typically approached this empirical anomaly in two ways. The first approach focuses on the misestimation of beta arising from the estimation method (Dimson, 1979; Scholes and Williams, 1977). The second approach focuses on anomaly-based explanations such as betting against beta (BAB) (Frazzini and Pedersen, 2014), the low-volatility effect (Blitz and

van Vliet, 2007; Blitz, van Vliet and Baltussen, 2019), lottery premia effects (Bali, Brown, Murray and Tang, 2017), and idiosyncratic volatility (Liu, Stambaugh and Yuan, 2018).

Using DTW, I extend the work of Dimson (1979) and Scholes and Williams (1977) and incorporate a dynamic asynchronicity adjustment in the estimation of beta, without imposing any strict assumptions on the lead–lag structure. The Dimson (1979) and Scholes and Williams (1977) beta estimates use leading and lagging market return variables as additional independent variables in the regression model used to estimate beta. However, both methods assume that the asynchronicity between the market and stock returns is static within the set of defined leading and lagging variables. The dynamics of the lead–lag, however, need not necessarily be fixed and can vary across time. DTW can more accurately align stock returns and market returns, allowing for a more flexible incorporation of non-synchronous trading effects into the estimation of beta.

By using DTW to account for the dynamic asynchronicity between stock and market returns, I recover a positive relation between the DTW-adjusted beta and expected returns. A stock's beta can be more accurately estimated using DTW, helping resolve the beta anomaly. Small stocks primarily drive the misestimation of beta. Smaller, less liquid stocks take longer to incorporate market-wide information, hence lagging the market. This influence of smaller stocks manifests with the largest difference between CAPM betas and DTW-adjusted betas occurring for smaller stocks. Despite the positive relation between DTW beta and expected returns, I cannot claim full resolution of the beta anomaly. From 2000, across all beta estimates, high beta stocks underperform low beta stocks, suggesting that other factors, in addition to non-synchronous trading, contribute to the beta anomaly. This result brings into question recent literature on the BAB phenomenon Frazzini and Pedersen (2014). The results I find are more aligned with the work of Novy-Marx and Velikov (2022), who call into question some of the exceedingly strong performance of the BAB factor.

The second empirical application uses DTW to study intraday dynamics in price leadership between two instruments. Established measures of price discovery, such as Hasbrouck's (1995) information share (IS) and Gonzalo and Granger's (1995) component share (CS), typically provide a summary statistic of price leadership between two instruments across an estimation window. These methods do not provide insight into the dynamics of the price discovery process within the estimation window. Ozturk, van der Wel and van Dijk (2017) propose a novel approach using flexible Fourier transformations to measure the intraday dynamics in IS by allowing

for time-varying volatility of the efficient price innovations and idiosyncratic noise. By applying DTW, an estimate of the lead–lag at each time-step in the estimation window is obtained, allowing a clearer insight into the intraday temporal dynamics between time series. Specifically, I explore intraday price discovery dynamics across global index futures.

The application of DTW uncovers a rich temporal behavior of the intraday lead–lag between global futures contracts, anchored around significant market operation events, such as the opening and closing of the underlying equity markets. Specifically, I measure the intraday lead–lag between the U.S. E-mini S&P 500 futures (E-mini) and U.K. FTSE 100 futures (FTSE 100), respectively traded on the Chicago Mercantile Exchange (CME) and the Intercontinental Exchange (ICE). These contracts represent two of the largest and most liquid equity markets, where I use the New York Stock Exchange (NYSE) and London Stock Exchange (LSE) trading hours as representative of.. The results show that the lead–lag structure between the two index futures has evolved. I document a compression in the average daily lead–lag toward zero from 2001 to 2010. As markets have become increasingly automated, the lead–lag between the two contracts is more likely to be restricted by the speed of information transmission. However, even in the modern highly electronic and liquid markets as in 2020, there is still a rich, dynamic intraday behavior between these futures contracts, highlighting the importance of developing price discovery measures that can quantify these dynamics.

Through these two empirical applications, I show that asynchronicity is highly dynamic and can affect inference in asset pricing, risk, and price discovery models. Using DTW to account for dynamic asynchronicity, new insights into empirical problems can be generated, and existing empirical challenges can be resolved. The use of DTW provides a pathway to the further study of existing and new issues where the asynchronicity inherent in financial time series obfuscates model estimation and inference. One such problem is in the emerging literature around the recent increase in co-movement in markets (e.g., from high-frequency trading (Malceniece, Malcenieks and Putniņš, 2019)). It is currently difficult to attribute this increase in co-movement to a genuine change in the systematic risk in markets. An alternative explanation proposes that the increase in co-movement is owing to a better alignment of returns in assets within markets that arises owing to the increased liquidity and trading activity afforded to stocks when added to an index (Barberis, Shleifer and Wurgler, 2005). Li, Yin and Zhao (2020) explore the effects of program trading on the co-movement of stocks and document evidence that stocks preferred by algorithmic traders exhibit excessive co-movement above what would be expected by fundamental drivers of return. This empirical problem is an example where a richer

insight into the nature of the effect of program trading on co-movement could be obtained by using DTW to account for the dynamic asynchronicity that is present in returns.

Finance literature on asynchronicity in time series has evolved as the speed of information transmission has increased. Earlier studies (Scholes and Williams, 1977; Dimson, 1979; Kawaller, Koch and Koch, 1987) typically use lagged variables in a regression framework to account for the intertemporal relation between financial time series. As trading in markets has become predominantly automated and the speed of trading has increased, regression frameworks to estimate lead–lag are often inadequate as the lead–lag relation has reduced from minutes to seconds to milliseconds. In the past 20 years, new techniques have emerged for measuring high-frequency lead–lag. Hayashi and Yoshida (2005) develop a covariance estimator for non-synchronous diffusion processes that accounts for temporal dynamics between time series. Dobrev and Schaumburg (2016) present a model-free method for estimating the lead–lag relation using a timing offset between trading activity to estimate which market is driving the lead–lag. The primary difference is that point-in-time estimates of the lead–lag relation can be obtained at every observation within the estimation window using the DTW approach proposed in this chapter. Point-in-time estimates allow for an examination of the time-varying dynamics in the lead–lag relation across the estimation window, which is often not possible with the covariance estimator of Hayashi and Yoshida (2005).

This work also contributes to the literature on synchronizing returns across markets with respect to different market closing times. The difference in market closing times must be accounted for when modeling assets that are traded in different markets. Otherwise, estimates that are derived from time series of these assets can produce misestimation of models. A significant body of literature proposes various methods that aim to synchronize returns across different markets using various statistical models (Burns, Engle and Mezrich, 1998; Martens and Poon, 2001; Audrino and Bühlmann, 2004; Scherer, 2013). DTW is a direct complement to these approaches and could also be applied to synchronize stock returns across different markets.

Ito and Sakemoto (2020) and Franses and Wiemann (2020) are perhaps the most closely related studies to this chapter, using DTW to study lead–lag relations in foreign exchange markets and business cycles, respectively. Ito and Sakemoto (2020) propose using DTW to measure high-frequency foreign exchange lead–lag patterns. Like this chapter which explores the dynamics of intraday price discovery, they document a change in the lead–lag patterns between currency pairs in response to important market announcements. I expand on their work by further exploring the

sensitivity of the optimal series alignment based on DTW to noise and volatility. I also provide further empirical applications on the use of DTW for studying asynchronicity in financial time series.

The rest of this chapter is structured as follows. Section 2.2 uses a simulation study to evaluate the use of DTW for measuring lead–lag patterns. Section 2.3 presents the results of applying DTW to account for asynchronicity when measuring a stock's beta. Section 2.4 presents results from using DTW to measure intraday price leadership dynamics between E-mini and FTSE 100 futures contracts. Section 2.5 then concludes the chapter.

2.2 Validating dynamic time warping

2.2.1 Simulation design

I use simulation to study and validate the effectiveness of the proposed DTW method for measuring lead–lag between financial time series. When applying DTW, the time series must share some commonality, as the DTW algorithm will find a relation where there is none. Thus, to validate DTW I create artificial lead–lag patterns between time series and then use DTW to recover these simulated patterns. I follow Putniņš (2013) and use a time-series model where two stocks share a common fundamental value. The fundamental value is assumed to follow a random walk:

$$m_t = m_{t-1} + u_t,\ u_t \sim \mathcal{N}(0, \sigma_u), \tag{2.1}$$

where m_t is the natural logarithm of the fundamental value at time t and u_t is a noise component. I define a time series p_i that tracks the fundamental value with a state-dependent time-shift of $\delta_{i,t}$ periods and noise $s_{i,t}$ as:

$$p_{i,t} = m_{t-\delta_{i,t}} + s_{i,t},\ s_{i,t} \sim \mathcal{N}(0, \sigma_{s_i}). \tag{2.2}$$

To test the efficacy of the DTW algorithm at capturing time-varying asynchronicity, three lead–lag scenarios representative of typical lead–lag behavior in financial markets are used: constant, gradual, oscillating, and, for robustness, a randomly switching lead–lag. The constant lead–lag scenario represents a pair of signals with a constant structural lag between them, for example, from trading on multiple venues where one venue dominates the price discovery process. The gradual lead–lag scenario represents a pair of signals where the lag between the two stocks decreases over the course of the trading day. This may occur from overnight information imbalances, that shrink as information is incorporated into the price

across the trading day. The oscillating lead–lag scenario represents a pair of signals where, over the trading day, the lag alternates between positive and negative. Such a situation may arise from the opening of other international markets and the impounding of new information across markets. The randomly switching lead–lag scenario represents a pair of signals where there is no predetermined dynamic asynchronicity. Thus, the lead–lag relation between the signals may switch randomly over the trading day. This scenario tests the robustness of the DTW algorithm in capturing any dynamic asynchronicity pattern.

To create these four scenarios, p_1 is set as the reference time series (i.e., $\delta_{1,t}$ is fixed at a constant value) and p_2 as the secondary time series. The state-dependent time-shift of p_2, $\delta_{2,t}$, is varied whilst $\delta_{1,t}$ is fixed to induce the desired lead–lag dynamics between the two time series. With $\delta_{1,t}, \delta_{2,t} \in \mathbb{N}$:

$$p_{1,t} = m_{t-\delta_{1,t}} + s_{1,t}, \tag{2.3}$$

$$p_{2,t} = m_{t-\delta_{2,t}} + s_{2,t}. \tag{2.4}$$

A constant lead–lag is induced by fixing $\delta_{2,t}$ to a constant value across the simulation. This produces a true lag of $\delta_{2,t} - \delta_{1,t}$. A gradual lead–lag is induced by setting $\delta_{2,0}$ to some initial value and then creating a sequence of evenly spaced lags $\delta_{2,t} = \left\lfloor \delta_{2,0} - \frac{\delta_2(N-t)}{N} \right\rceil$ for $t = 1, \ldots, N$ where N is the cardinality of $\{p_{2,t}\}$. An oscillating lead–lag takes the functional form $\delta_{2,t} = \lfloor A \sin(2\pi x_t) \rceil$ where $A \in \mathbb{N}$ is a scaling factor, and x_t is taken from an evenly spaced grid of values, $x = \{x_0, x_1, \ldots, x_t\} = \left\{0, \frac{1}{N}, \ldots, 1\right\}$. This induces a sinusoidal wave of period 2π and amplitude of one. A scaling factor, A, is used to control the peak-to-trough range of p_2. A randomly switching lead–lag is induced by randomly sampling from $\mathcal{N}(0, 1)$ and if the sampled value is greater than 3.7, a new value of $\delta_{2,t}$ is drawn from $\mathbb{N}(20)$.

2.2.2 Dynamic time warping

There are several variants of DTW which have been proposed. These variants aim to address several shortcomings with the base DTW algorithm, such as the pathological alignment problem. In this problem, the DTW algorithm aligns a single point from one time-series to a large sub-sequence of points from the other time-series. Constrained DTW is a common approach to mitigating the pathological alignment problem, whereby the warping path is restricted by applying a windowing function (e.g., Sakoe and Chiba, 1978). Other variants, such as the derivative DTW (Keogh and Pazzani, 2002), weighted DTW (Jeong, Jeong and Omitaomu, 2011), and shape DTW (Zhang, Tang and Duan, 2015) all present innovations on the base DTW algorithm which aim to improve the alignment quality. There is no strong consensus

in the literature on which DTW algorithm is optimal, and varied performance is observed across different benchmarking tasks(Lahreche and Boucheham, 2021).

I use the generalized version of derivative DTW combined with a Sakoe-Chiba constraint throughout. The selection of this method acknowledges that a key limitation of DTW is that the algorithm can map time-series points together which occur very far away in time. In financial settings, such mappings are likely noise and thus a constrained DTW approach helps to alleviate this problem. I present a DTW algorithm where the two time series can take differing lengths, N and M, such that $p_{1,t} = (p_{1,1}, p_{1,2}, \dots p_{1,N})$ and $p_{2,t} = (p_{2,1}, p_{2,2}, \dots, p_{2,M})$. An $N \times M$ cost matrix $C \in R^{N \times M}$ is computed, where each element in row n and column m corresponds to the distance between the elements $(p_{1,n}, p_{2,m})$. Typical distance measures used include the Euclidean distance and Manhattan distance, I use the Euclidean distance throughout the rest of this chapter.

The objective of the DTW algorithm is to find the optimal alignment between p_1 and p_2, where optimality is the alignment that has the minimum total cost. A warping path $Z = (z_1, z_2, ..., z_K)$ with $z_k = (n_k, m_k)$ is defined as the set of matrix elements that define a mapping between p_1 and p_2, where the following conditions are satisfied:

- Boundary condition: $z_1 = (1, 1)$ and $z_K = (N, M)$
- Monotonicity condition: $n_1 \leq n_2 \leq \ldots \leq n_K$ and $m_1 \leq m_2 \leq \ldots \leq m_K$
- Step-size condition: $z_{l+1} - z_l \in \{(1, 1), (1, 0), (0, 1)\}$

I note that these conditions are specific to the selected DTW algorithm, and can be relaxed or altered. For example, the step-size condition need not always be one unit, and could instead be extended. The path that has the minimum total cost from all possible warping paths is denoted as the optimal warping path. Dynamic programming methods are used to calculate an accumulated cost matrix, D. Each entry in the accumulated cost matrix is defined as the local cost measure in the current C matrix cell, $c(x_n, y_m)$, plus the minimum of the accumulated cost measure in adjacent cells from D:

$$D(n, m) = \min(D(n-1, m-1),\ D(n-1, m),\ D(n, m-1)) + c(x_n, y_m). \quad (2.5)$$

The optimal warping path Z^* is calculated by stepping in reverse index order through the accumulated cost matrix D, following the algorithm:

$$w_{k-1} = \begin{cases} (1,\, m-1) & \text{if } n = 1, \\ (n-1,\, 1) & \text{if } m = 1, \\ \arg\min(D(n-1, m-1),\, D(n-1, m),\, D(n, m-1)) & \text{otherwise.} \end{cases}$$

Given the empirical settings I intend to test, I impose and test a set of global constraints. The classic Sakoe-Chiba (1978) band that runs along the main diagonal and has a fixed width $W \in \mathbb{N}$ is used. This is a global constraint, that implies that an element $p_{1,n}$ can only be aligned to some value $p_{2,m}$ where $m \in \left[\frac{m-W}{n-W}(N-W),\, \frac{m-W}{n-W}(N+W)\right] \cap [1:m]$. The boundary condition is also loosened, allowing $z_1 \in (1,1) : (1+\psi,\, 1+\psi)$ and $z_k \in (n-\psi,\, m-\psi) : (n,\, m)$ where $\psi \in \mathbb{N}$. The boundary condition is loosened so that the start and end points of the time series are not required to align perfectly at the start and end of the trading day, as is often the case for high-frequency financial time-series data.

The optimal warping path Z^* is used as a measure of the asynchronicity between p_1 and p_2 at each time-step t. The optimal warping path will have at minimum a length of $\max(N, M)$, this can result in there being more lags in the optimal warping path than time-steps. This occurs due to index duplication where the lag does not change. To map from warp-time back to clock-time, all index pairs $z_k = (n_k, m_k)$ are taken and for each duplicated value of m_k I take the first occurrence of m_k and map this to time-step k. The asynchronicity at each time-step is then measured as:

$$\delta_k = n_k - m_k. \tag{2.6}$$

A measure of performance is crucial in assessing the effectiveness of DTW in recovering the true lag in the simulated time series. As we seek to measure the difference of the measured lag versus the true lag we impose into the time series, for the true lag, δ_i, and the lag recovered by DTW, $\hat{\delta}_i$, I define:

$$\text{MAE} = \frac{\frac{1}{N}\sum_{i=1}^{N}\left|\delta_i - \hat{\delta}_i\right|}{\frac{1}{N}\sum_{i=1}^{N}|\delta_i|} \text{ if } \frac{1}{N}\sum_{i=1}^{N}|\delta_i| \neq 0. \tag{2.7}$$

An alternative approach could be to use an information criterion approach, as what is performed when selecting the order of a vector autoregression model, to measure how well the DTW estimated lags compare to the true lags. However, such an

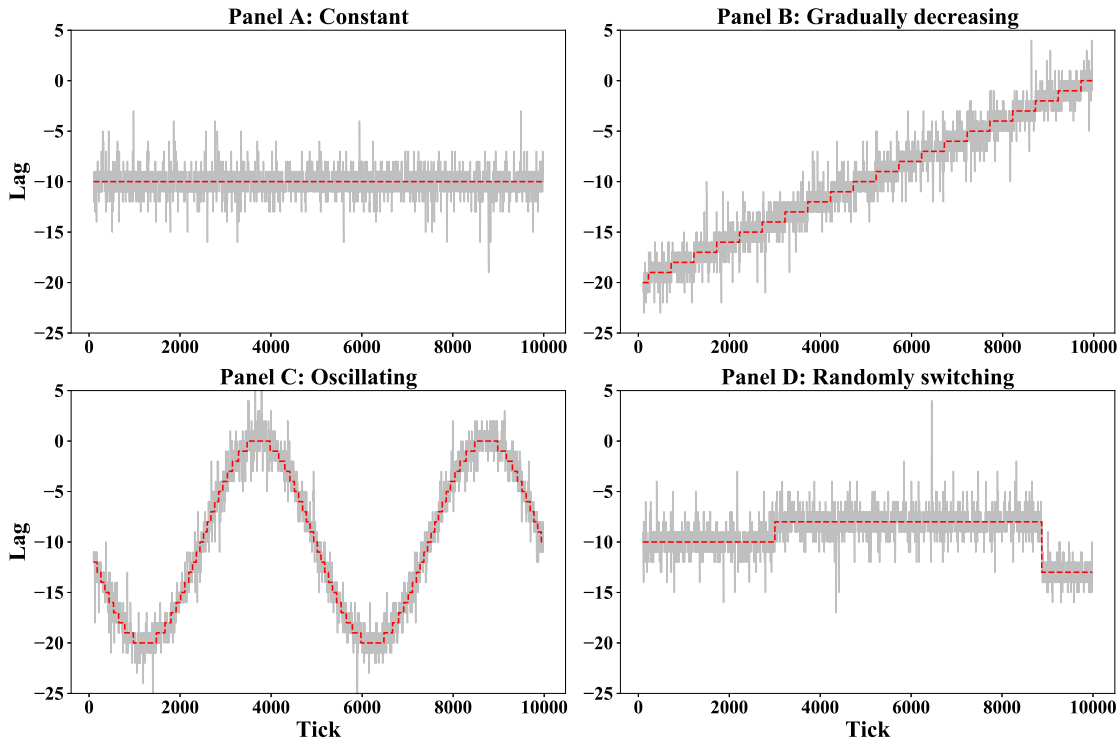

Figure 2.1: DTW-estimated lead–lag in four simulated scenarios
This figure presents the lead–lag estimated using DTW in four scenarios over a single simulation. In each scenario, an artificial lead–lag structure is induced between two simulated time series that share a common fundamental value. DTW is applied to the two time series to estimate the artificial lead–lag. The solid gray line is the lead–lag estimated using DTW. The dashed red line is the induced lead–lag.

approach is likely not appropriate in the proposed setting, given that I seek to allow the DTW algorithm to flexibly select the optimal mappings without imposing any structural assumptions.

2.2.3 Simulation results

Figure 2.1 demonstrates an example of the induced lead–lag structures (dashed red lines) between two simulated time series and the lead–lag recovered from the DTW algorithm (solid gray lines). In each of the four scenarios, the DTW algorithm successfully recovers the simulated lead–lag pattern.

Table 2.1 presents several metrics from the four baseline simulations. Across all four cases, I find consistent results. In the baseline scenarios, a lag of negative ten ticks (i.e., the difference in speed between p_1 and p_2 is ten ticks) is applied and DTW recovers the induced lag with a MAE between 5.83% and 6.50%. The error in DTW's ability to recover the lag is a combination of bias in the underlying DTW algorithm and noise in the two time series. The bias inherent in the DTW algorithm

Table 2.1: Baseline DTW estimation results for four lead–lag scenarios
This table presents the results when measuring lead–lag using DTW in four scenarios over a single simulation. Following Eq. (2.1) and Eq. (2.2) defined in Section 2.2.1, two time series (p_1 and p_2) that share a common fundamental value (u_t) are simulated. Both time series track the common fundamental value with some time-delay ($\delta_{i,t}$) and noise ($s_{i,t}$). The parameters used in the simulation are $N = 10,000$, $u_t = 1$, $s_{1,t} = s_{2,t} = 0.5$, and $W = 60$. Each simulation aims to produce a true lag of ten (i.e., p_2 lags p_1 by ten units on average across the simulation). In the constant lead–lag scenario $\delta_{1,t} = 5$ and $\delta_{2,t} = 15$. In the gradual lead–lag scenario $\delta_{1,t} = 5$ and the initial $\delta_{2,0} = 30$, $\delta_{2,t}$ is then reduced as t increases. In the oscillating scenario the lead–lag takes the functional form of $\delta_{2,t} = 10\sin(2\pi x_t)$ where $x_t = \{1, 2, ..., 1000\}$, and $\delta_{1,t} = 5$. In the switching lead–lag scenario the initial values are fixed at $\delta_{1,t} = 5$ and $\delta_{2,0} = 15$, and then $\delta_{2,t}$ is randomly switched between 0 and 20.

	Lead–lag Scenario			
	Constant	Gradual	Oscillating	Switching
Average of true lag	-10.00	-10.04	-10.00	-9.16
Average of DTW lag	-10.13	-9.97	-10.07	-9.21
DTW distance	0.84	0.85	0.84	0.86
MAE (absolute)	0.58	0.58	0.59	0.60
MAE (proportional to true lag (%))	5.86	5.83	5.86	6.50
STD. of true lag - DTW lag	0.94	0.89	0.95	0.95
Pearson correlation between true lag and DTW lag[1]	-	0.98	0.99	0.80

is a function of how the algorithm stitches the two time series together. DTW, like many algorithms, is prone to biases. The DTW algorithm tries to minimize the total cumulative distance between the two time series, however, this can result in localized stitching errors due to selection choices that the algorithm must make. I find a directional bias by switching which time series is the primary time series in the DTW algorithm. This bias is corrected in subsequent analyses by running a two-step DTW. To correct for this bias, the primary time series given to the DTW algorithm is alternated, and then the difference of the final results divided by two is used as the DTW lead–lag estimate. Table 2.13 in Appendix 2.5 examines the characteristics of this bias, and shows that after adjusting for the bias, the DTW algorithm successfully recovers the induced true lead–lag with minimal error.

In addition to standalone simulations, a bootstrapping study of the simulation parameters is used to measure the sensitivity of the DTW algorithm in recovering the induced lead–lag patterns. The MAE is measured over 1,000 pairs of simulated time series under different parameter combinations. In Figure 2.2 the time-delay ($\delta_{2,t}$) and noise (σ_{s_2}) of the second time series is varied, measuring the accuracy of DTW when the absolute level of the lag and the ratio of noise between the two time series changes. The performance of DTW under these conditions is relevant

[1]Pearson's correlation is undefined for the constant simulation as there is no change in the true lag across the simulation.

to applications where the volatility of the assets being studied is different. $\delta_{2,t}$ is varied between zero and ten, and σ_{s_2} is varied between zero and two in each of the four scenarios. In these simulations, the time-delay of the first price-series is fixed ($\delta_{1,t} = 5$), thus when $\delta_{2,t} = 5$, the induced lag is zero. When the induced lag is zero, MAE is highest, a result of the directional bias in the DTW algorithm, that produces a larger MAE when the true lag is close to zero. This effect is also present in the gradual lag and switching lag scenarios. In the gradual lag scenario, this increase in MAE occurs when $\delta_{2,t} = 6$ due to how the gradual lag structure is induced. In the oscillating scenario, the results are primarily a function of σ_{s_2}, as the MAE is relatively constant for different levels of $\delta_{2,t}$. The biased MAE does not manifest in the oscillating scenario due to the symmetric nature of the DTW bias, The oscillating lag scenario is constructed to have an area under of the curve of zero (i.e., the lag structure is symmetric), that cancels out the bias and results in an approximately linear relation between MAE and σ_{s_2}. Overall, I conclude that error in the DTW algorithm is predominantly a function of the relative noise in the time series. As the level of noise in one time series increases relative to another time series, the error will also increase. It is important to consider this when applying DTW, but it does not preclude the use of DTW when studying lead–lag in highly volatile financial time series.

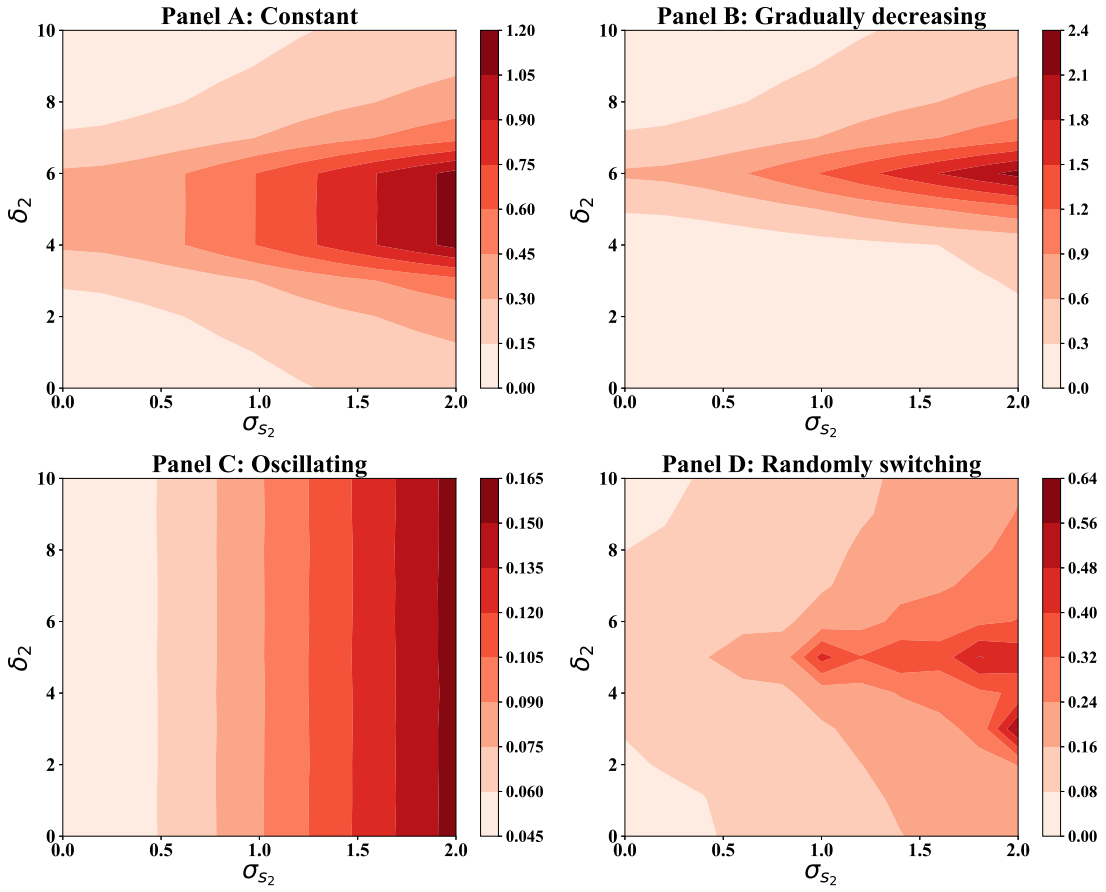

Figure 2.2: Effect of varying time-delay and noise on the accuracy of the DTW lead–lag estimation
This figure presents the MAE between the artificially induced lead–lag and the DTW-estimated lead–lag when varying the noise (σ_{s_2}) and time-delay ($\delta_{2,t}$) of the second time series in four scenarios described in Section 2.2.1. The MAE is the difference between the DTW-estimated lag and artificially induced lag at each tick of the simulation, proportional to the average of the induced lead–lag across the simulation. I use a grid of ten values for σ_{s_2} and $\delta_{2,t}$. σ_{s_2} is varied between zero and two at increments of 0.2. $\delta_{2,t}$ is varied between zero and ten at increments of one unit. For each possible pair of parameters, 1,000 simulations are run with input parameters of $N = 10,000$, $\sigma_u = 1$, $\delta_{1,t} = 5$, $\sigma_{s_1} = 1$, $W = 60$, and $\psi = 60$. Each panel corresponds to one lead–lag scenario and the average MAE across the 1,000 simulations.

Having established that the relative level of noise between the two time series is a driver of error in the DTW algorithm, I explore the effect of three additional parameterizations on the accuracy of the DTW algorithm. Figure 2.3 presents three different parameter pair combinations for the gradually decreasing lag scenario. Panel A presents the MAE when varying the noise in the two time series, σ_{s_1} and σ_{s_2}. The MAE is linearly increasing in both σ_{s_1} and σ_{s_2}. This result is expected, σ_{s_1} and σ_{s_2} control the level of noise in the underlying time series, and as more noise is induced, the DTW algorithm makes more errors when stitching the two time

series together. As I aim to apply DTW to measure lead–lag at high-frequency, the algorithm must be stable for long time series. To test this, MAE is measured for varying levels of the time-series length (N) and the volatility of the fundamental value (σ_u). In panel B the MAE is reducing as $\log(N)$ increases and constant in σ_u. σ_u is common to both time series and thus the only variability that is induced is in the noise of the two time series relative to the fundamental value, not in the fundamental value itself. In panel C, we observe a non-linear relation between σ_{s_2} and σ_u. At lower levels of σ_{s_2}, the MAE is higher. This result is likely owing to the low signal-to-noise ratio between the fundamental value and the observed time series. Ultimately, if the fundamental value is highly volatile, the DTW algorithm will have a higher error level. The DTW algorithm is more successful at recovering the lead–lag when the two time series have a similar level of noise. The behavior of DTW under different noise levels has implications for the usage of DTW in price discovery applications, suggesting that the DTW algorithm would perform better when the volatility levels in the two time series are similar.

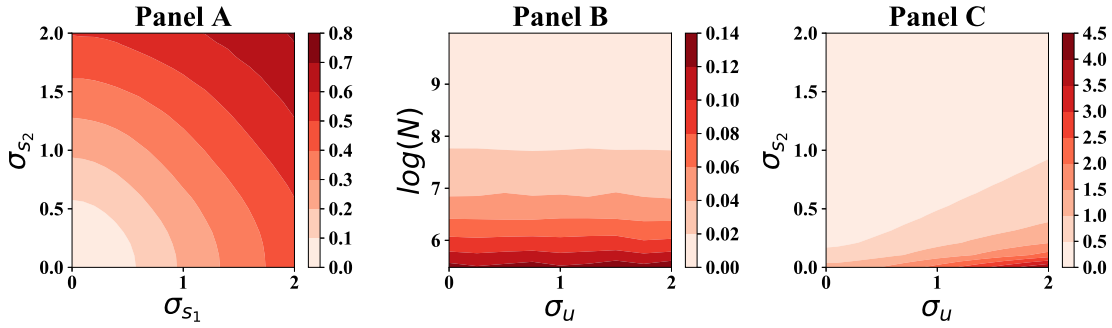

Figure 2.3: Effect of varying simulation parameters on the accuracy of the DTW lead–lag estimation
This figure presents the MAE generated from varying three different parameter pairs when generating the time series, and then estimating the lead–lag using DTW. In panel A: $\sigma_{s_1}$and σ_{s_2} are varied between zero and two, using 0.2 increments. In panel B: N is set as the natural logarithm of 240, 3,600, 7,200, 14,400, and 21,600 and σ_u is varied between zero and two. using 0.2 increments. In panel C: σ_u and σ_{s_2} are varied between zero and two, using increments of 0.2. The panels show the bootstrapped average MAE between the artificially induced lead–lag and the DTW recovered lead–lag. For each parameter combination, 1,000 simulations are run. In panel A: $\sigma_u = 1$, $\delta_{1,t} = 5$, $\delta_{2,t} = 10$, $W = 60$, $\psi = 6$, and $N = 10,000$. In panel B: $\delta_{1,t} = 5$, $\delta_{2,t} = 10$, $\sigma_{s_1} = 0.1$, $\sigma_{s_2} = 0.1$, $W = 60$, and $\psi = 60$. In panel C: $\delta_{1,t} = 5$, $\delta_{2,t} = 10$, $\sigma_{s_1} = 0.1$, $W = 60$, $\psi = 60$, and $N = 10,000$.

2.3 A better beta

The CAPM predicts that stocks with high betas to the market will outperform stocks with low betas (Black, Jensen and Scholes, 1972; Fama and MacBeth,

1973). However, several empirical studies (Reinganum, 1981; Lakonishok and Shapiro, 1986; Fama and French, 1992) provide evidence that stocks with low betas outperform stocks with high betas. This result is commonly referred to as the "beta anomaly" or "low beta" effect. As one of the key empirical challenges to the CAPM, extensive literature attempts to explain the existence of the beta anomaly. There are typically two approaches to explanation. The first stream posits that standard methods misestimate beta under the CAPM, and they present alternative estimation techniques, such as the approaches of Dimson (1979) and Scholes and Williams (1977). The second stream focuses on anomaly-based explanations and pose that controlling for other cross-sectional stock characteristics resolves the beta anomaly. Liu et al. (2018) propose a resolution by controlling for idiosyncratic volatility while Bali et al. (2017) argue that investor demand for lottery-like stocks is a main driver of the beta anomaly. Although these approaches control for and render the beta anomaly insignificant under a Fama-Macbeth regression framework, they do not fully recover the expected relation that the CAPM predicts, that stocks with high beta outperform stocks with low beta. I contribute to the literature by providing an alternative method to estimate beta. Using DTW, the dynamic asynchronicity between stock returns and market returns can be flexibility incorporated into the measurement of beta. This approach recovers a weak positive relation between stock betas and future returns, that is robust to controlling for well-established asset pricing factors, including idiosyncratic volatility and the price lottery effect.

When using DTW to estimate beta, an important consideration is whether there exists asynchronicity between stock returns and market returns. If there is little to no asynchronicity, the application of DTW will predominantly capture noise and result in worse estimates of beta. Figure 2.4 depicts the proportion of non-contemporaneous correlation between the market return and all U.S. stocks in the Center for Research in Security Prices (CRSP) sample, using one-day forward and one-day lagged stock returns, relative to the correlation using one-day forward, contemporaneous, and one-day lagged stock returns. This figure proxies the degree of asynchronicity in U.S. equity markets. The dynamic nature of asynchronicity is apparent in Figure 2.4, with significant time variation in the proportion of non-contemporaneous correlation. The time-varying nature of asynchronicity is present across all capitalization levels of stocks. Empirical models have traditionally addressed this non-contemporaneous correlation by leading or lagging market return variables in a regression framework. This approach imposes a pre-specified and fixed structure around the nature of the asynchronicity, and thus the models do not

account for dynamics in the asynchronicity. Using DTW to estimate beta removes this limiting assumption.

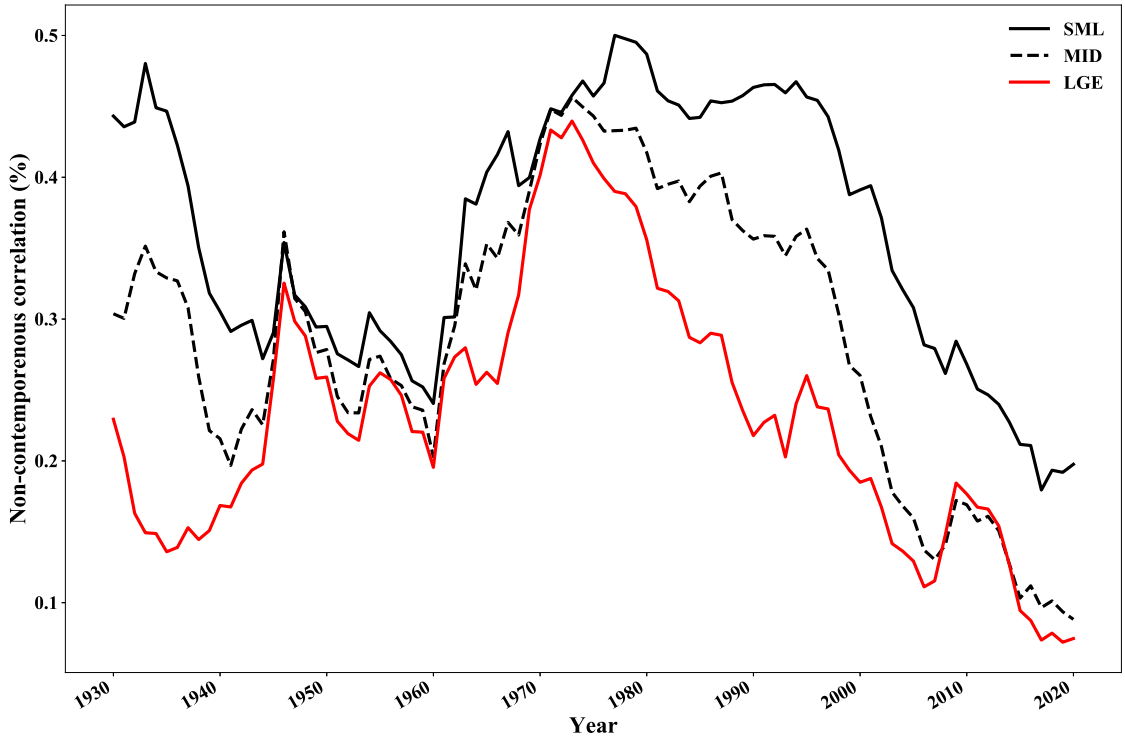

Figure 2.4: Non-contemporaneous correlation in the CRSP sample
This figure presents the ratio of the one-day forward correlation and one-day lagged correlation between CRSP stock returns and market returns to the sum of the one-day forward correlation, contemporaneous correlation and one-day lagged correlation. At the end of each month between June 1931 and December 2019, the Pearson correlation is calculated between stock returns and market returns, as $Corr(r_{m,t}, r_{i,t+s})$ with $s = (-1, 0, +1)$ using the previous five years of daily data. The non-contemporaneous correlation proportion is calculated:

$$\frac{|Corr(r_{m,t}, r_{i,t-1})| + |Corr(r_{m,t}, r_{i,t+1})|}{|Corr(r_{m,t}, r_{i,t-1})| + |Corr(r_{m,t}, r_{i,t})| + |Corr(r_{m,t}, r_{i,t+1})|},$$

where $r_{m,t}$ is the market return for day t and $r_{i,t+s}$ is the return of stock i at day $t+s$. The universe is split into three size buckets, SML/MID/LGE, based on 30%/70% market capitalization cutoffs of stocks traded on the NYSE. This figure presents the cross-sectional median non-contemporaneous correlation proportion within the three size buckets at the end of each month.

In addition to studying non-contemporaneous correlations, DTW can directly measure the daily lead–lag between stock returns and market returns. Table 2.2 presents the estimated daily lead–lag between stock returns and market returns for the standard U.S. CRSP sample. The results are intuitive, particularly for small stocks. Small stocks have evolved from lagging the market by an average of 0.81 days between 1950 and 1970 to exhibiting almost no lag between 1990 and

Table 2.2: DTW-estimated lead–lag between CRSP stocks and the market
This table presents the time-series average DTW-estimated lead–lag between daily U.S. stock returns and market returns. The sample consists of U.S. common ordinary share stocks traded on the NYSE, NASDAQ, and AMEX, starting in January 1950, and ending in December 2019. The sample is split into SML/MID/LGE stocks based on 30%/70% market capitalization cutoffs of stocks traded on the NYSE. The sample is also split into three time buckets, 1950–1970, 1970–1990, and 1990–2019. This table presents the average and standard deviation of the daily cross-sectional median lead–lag within each time bucket and size bucket.

	SML	MID	LGE
1950–1970	0.81 ± 2.07	-0.04 ± 1.35	-0.75 ± 0.95
1970–1990	0.58 ± 1.84	0.09 ± 1.31	-0.60 ± 1.09
1990–2019	0.07 ± 1.51	-0.17 ± 1.37	-0.51 ± 0.99

2019, alongside a reduction in the dispersion of lead–lag across small stocks. A key finding is that the lead of large stocks over the market has not changed significantly over the sample period. This persistent leading of large stocks suggests that the asynchronicity that the DTW method accounts for is not typically present in large stocks.

One possible explanation for the beta anomaly is that small, illiquid stocks are slower to incorporate market-wide information. Using contemporaneous market returns to estimate beta thus results in a downward biased beta. Several methods have been developed to address this non-synchronous trading effect. For instance, Dimson (1979) and Scholes and Williams (1977) combine contemporaneous market returns with pre-specified forward and lagged market returns when estimating beta. In contrast, I use DTW to align the stock return series with the market return series, thereby accounting for dynamic asynchronicity. The DTW-aligned returns series is then used as an input to estimate beta in a regression framework.

2.3.1 Data and method

Daily and monthly stock data are obtained from CRSP. I take all U.S. stocks traded on all exchanges that are classified as ordinary common shares (SHRCD 10 or 11) with a stock price (PRC) greater than $5, and adjust returns for delisting bias as per Shumway (1997) and Shumway and Warther (1999). Daily and monthly market returns, high-minus-low (HML), small-minus-big (SMB), robust-minus-weak (RMW), conservative-minus-aggressive (CMA), and up-minus-down (UMD) factor returns, and risk-free (one-month Treasury Bill) rates are from Kenneth French's data library. Daily FMAX factor returns are from Turan Bali's website (Bali et al., 2017). Balance sheet information used to calculate the book-to-market ratio is

sourced from Compustat. The sample covers the months t from July 1927 through November 2019 and month $t+1$ returns from August 1927 through December 2019.

Using DTW, the best alignment of stock returns and market returns is systematically selected. This alignment is used to remove the effect of asynchronicity on the ex-post estimation of beta. I calculate several stock characteristics as part of the DTW-estimation process: raw DTW beta (β_{DTWR}), bootstrapped DTW beta ($\hat{\beta}_{DTW}$), DTW beta (β_{DTW}), and the bootstrapped DTW t-statistic (β_{DTWT}). β_{DTWR} is estimated directly from the aligned market returns and stock returns. A bootstrapping approach is applied to the estimation to address the inherent bias in the DTW method, as described in Section 2.2.3. The stock returns series is randomly permuted, DTW is used to align the permuted stock returns and market returns, and then $\hat{\beta}_{DTW}$ is re-estimated. This process produces a de-biased DTW beta which is defined as the difference between the β_{DTWR} and $\hat{\beta}_{DTW}$. In addition to the DTW estimates of beta, I also calculate the bootstrapped t-statistic associated with estimating $\hat{\beta}_{DTW}$. The full calculation details can be found in Appendix 2.5.

Alongside various measures of beta, several control variables are defined. MAX is the largest return over the prior month. MIN is the smallest return over the prior month. SIZE is the natural logarithm of the market capitalization. ILLIQ is Amihud's illiquidity measure. BM is the book-to-market ratio. MOM is the return from months $t-11$ to $t-1$. REV is the prior month's return (short-term reversal). IVOL is the idiosyncratic volatility from a 12-month regression of daily returns. TSKEW is the total skewness, ISKEW is the idiosyncratic skewness, and SSKEW is the systematic skewness calculated from daily returns over the past 12 months.

2.3.2 Results and discussion

Table 2.3: Summary statistics
This table presents summary statistics for DTW-adjusted beta measures and a set of control variables. For each month between July 1927 and November 2019, I calculate the cross-sectional average (Mean), standard deviation (SD), skewness (Skew), kurtosis (Kurt), number of observations (Nobs), minimum (Min), maximum (Max), median ($q_{0.5}$) and 5th ($q_{0.05}$), 25th ($q_{0.25}$), 75th ($q_{0.75}$), 95th ($q_{0.95}$) percentiles. The table presents the time-series average of these cross-sectional statistics. β_{DTW} is the DTW-adjusted beta after adjusting for bias in the DTW estimator using bootstrapping. β_{DTWT} is the t-statistic associated with the bootstrapping used to calculate β_{DTW}. β_{DTWR} is the raw DTW-adjusted beta. β is the beta estimated from a regression of daily excess stock returns on excess market returns using 12 months of daily returns. β_{DIM} is the beta estimated using the Dimson regression adjustment using 12 months of daily returns. β_{SW} is the beta estimated using the Scholes-Williams regression adjustment using 12 months of daily returns. MAX is the largest return over the prior month. MIN is the smallest return over the prior month. SIZE is the natural logarithm of the market capitalization. ILLIQ is Amihud's illiquidity measure. BM is the book-to-market ratio. MOM is the return from months $t-11$ to $t-1$. REV is the prior month's return (short-term reversal). IVOL is the idiosyncratic volatility from a 12-month regression of daily returns. TSKEW is the total skewness, ISKEW is the idiosyncratic skewness, and SSKEW is the systematic skewness calculated from daily returns over the past 12 months. Full variable definitions are provided in Appendix 2.5.

	Mean	SD	Skew	Kurt	Nobs	Min	Max	$q_{0.5}$	$q_{0.05}$	$q_{0.25}$	$q_{0.75}$	$q_{0.95}$
β_{DTW}	0.39	0.39	0.39	8.82	2276	-1.76	2.68	0.35	-0.16	0.14	0.60	1.06
β_{DTWT}	22.33	4.12	-0.02	0.42	2276	7.78	36.55	22.24	15.83	19.48	25.12	29.17
β_{DTWR}	1.08	0.57	1.13	5.48	2276	-0.51	4.85	0.99	0.34	0.68	1.40	2.13
β	0.92	0.57	0.52	0.76	2276	-0.86	3.27	0.86	0.12	0.50	1.28	1.95
β_{DIM}	1.07	0.76	0.45	3.42	2273	-2.63	5.35	1.01	0.01	0.57	1.51	2.38
β_{SW}	0.86	0.52	0.49	2.70	2276	-1.19	3.51	0.82	0.11	0.50	1.18	1.78
MAX	0.03	0.02	2.74	32.74	2024	0.00	0.23	0.03	0.01	0.02	0.04	0.06
MIN	-0.03	0.01	-1.43	5.35	2024	-0.13	0.00	-0.03	-0.05	-0.03	-0.02	-0.01
SIZE	11.37	1.61	0.45	0.04	2420	7.26	17.37	11.21	8.98	10.20	12.40	14.24
ILLIQ	1.47	3.92	10.18	243.52	1845	0.00	71.70	0.52	0.04	0.17	1.45	5.63
BM	0.78	0.61	5.24	101.83	1736	0.03	12.06	0.67	0.17	0.40	1.01	1.71
MOM	0.19	0.47	3.69	47.92	2251	-0.70	7.20	0.11	-0.31	-0.07	0.33	0.92
REV	1.70	11.03	2.28	38.36	2409	-41.53	135.30	0.70	-12.83	-4.31	6.36	19.08
IVOL	0.33	0.22	3.97	66.76	2412	0.02	3.49	0.29	0.12	0.20	0.41	0.70
TSKEW	0.46	1.10	2.22	24.02	2276	-6.55	10.97	0.34	-0.75	0.00	0.77	2.04
ISKEW	0.53	1.15	1.97	21.67	2276	-6.78	11.02	0.42	-0.81	0.05	0.87	2.22
SSKEW	-3.57	16.43	0.33	14.58	2276	-99.84	108.77	-3.15	-29.97	-12.22	5.32	0.39

Table 2.3 presents the time-series average of the cross-sectional summary statistics of the DTW beta factors, the standard CAPM beta estimate (β), Dimson beta (β_{DIM}), Scholes-Williams beta (β_{SW}), and a set of control factors. The full details on factor construction can be found in Appendix 2.5. β_{DTWR} has a mean larger than the three standard beta measures, suggesting a downward beta bias that is corrected when accounting for dynamic asynchronicity in the estimation of beta. However, there is a bias in the DTW algorithm, that can result in spurious estimates of beta. Although β_{DTWR} may provide a fair representation of beta, there is a strong relation with

Table 2.4: Historical correlations of asynchronicity-adjusted betas with the CAPM beta

This table presents the average Spearman rank correlation of $\beta_{DTW}, \beta_{DIM}, \beta_{SW}$ with β across different sample windows. β is the CAPM beta estimated from a regression of daily excess stock returns on excess market returns. β_{DIM} is the beta estimated using the Dimson regression adjustment. β_{SW} is the beta estimated using the Scholes-Williams regression adjustment. β_{DTW} is the beta estimated using DTW to align stock and market returns. A rolling regression is estimated at the end of each month using the last 12 months of daily returns data.

	All			Large			Small			Micro		
	β_{DIM}	β_{SW}	β_{DTW}	β_{DIM}	β_{SW}	β_{DTW}	β_{DIM}	β_{SW}	β_{DTW}	β_{DIM}	β_{SW}	β_{DTW}
July 1927–December 2019	0.66	0.87	0.53	0.69	0.89	0.66	0.66	0.87	0.63	0.60	0.82	0.50
July 1927–December 1999	0.66	0.87	0.54	0.68	0.89	0.65	0.67	0.88	0.64	0.60	0.82	0.52
July 1927–June 1963	0.68	0.88	0.59	0.73	0.90	0.68	0.68	0.88	0.62	0.60	0.83	0.52
July 1963–December 1999	0.64	0.86	0.49	0.64	0.88	0.62	0.67	0.88	0.65	0.60	0.81	0.52
January 2000–December 2019	0.65	0.87	0.51	0.70	0.90	0.69	0.62	0.84	0.61	0.62	0.85	0.45

idiosyncratic volatility, and increases in the level of beta, as measured by β_{DTWR}, coincide with increases in idiosyncratic volatility. Nonetheless, β_{DTWR} has practical applications from a risk management perspective, as it offers an innovative way to adjust for sluggishness in incorporating market-wide information into stock prices. Importantly, this finding supports Liu et al. (2018), who also examine the relation between beta and idiosyncratic volatility. Owing to the link between β_{DTWR} and idiosyncratic volatility, bootstrapping is used to correct for bias in the estimation of DTW beta, as described in Eq. (2.14).

Table 2.4 presents the average Spearman rank correlation between stock level estimates of β and β_{Sw}, β_{DIM}, and β_{DTW}. Using NYSE market capitalization breakpoints, the sample is divided into micro (bottom 20%), small (middle 30%), and large (top 50%). The correlation of β_{Sw}, β_{DIM}, and β_{DTW} with β is estimated across different time windows. Across all time horizons and size categories, the correlation with the β is relatively stable for the three asynchronicity-adjusted beta measures. The primary differences are the lower average correlations for micro stocks for each measure and β_{DTW} having a lower correlation with β than β_{DIM} and β_{SW}. This finding is most pronounced for micro stocks, suggesting that β_{DTW} is most different to β for micro stocks. This difference supports the notion that β_{DTW} has the most impact on beta estimation for micro stocks that lag the market and where the downward beta bias is larger for these stocks.

Table 2.5 shows the time-series average of the monthly median characteristic values for each decile portfolio produced by sorting from low to high β_{DTW}. A "U-shaped" pattern is observed across several characteristics, where portfolios one and ten have high values, and portfolios two to nine have relatively lower values. This pattern is particularly prominent for ILLIQ, both portfolio one and portfolio ten are comprised

Table 2.5: Portfolio characteristics of stocks sorted by DTW beta
Each month from July 1927 to November 2019, decile portfolios are formed by sorting stocks based on DTW-adjusted beta (β_{DTW}) over the previous month. This table presents the time-series average of the monthly median of characteristics within each decile portfolio.

Decile	β_{DTW}	β_{DTWR}	β	β_{DIM}	β_{SW}	BM	MAX	MIN	IVOL	ILLIQ	ISKEW	MOM	PRC	REV	SIZE	SKEW	TSKEW
Low β_{DTW}	-0.16	0.46	0.61	0.65	0.58	0.69	0.03	-0.03	0.31	0.60	0.50	0.10	13.41	0.96	10.81	-3.88	0.45
2	0.03	0.55	0.57	0.66	0.55	0.73	0.02	-0.02	0.24	0.38	0.38	0.10	18.68	0.74	11.28	-2.71	0.33
3	0.14	0.65	0.61	0.73	0.60	0.73	0.02	-0.02	0.24	0.35	0.36	0.11	20.84	0.74	11.45	-2.61	0.30
4	0.23	0.76	0.68	0.81	0.66	0.71	0.02	-0.02	0.24	0.38	0.36	0.12	21.99	0.73	11.52	-2.70	0.30
5	0.31	0.87	0.76	0.91	0.73	0.69	0.02	-0.02	0.25	0.42	0.37	0.12	22.56	0.76	11.53	-2.82	0.30
6	0.39	0.99	0.85	1.01	0.81	0.68	0.03	-0.02	0.26	0.50	0.39	0.12	22.50	0.77	11.50	-3.04	0.31
7	0.49	1.13	0.94	1.12	0.89	0.66	0.03	-0.02	0.28	0.55	0.40	0.12	21.66	0.71	11.44	-3.20	0.32
8	0.60	1.30	1.06	1.24	0.99	0.64	0.03	-0.03	0.30	0.66	0.43	0.12	20.38	0.78	11.33	-3.50	0.34
9	0.76	1.55	1.20	1.40	1.12	0.62	0.03	-0.03	0.33	0.96	0.46	0.12	18.04	0.73	11.14	-3.68	0.37
High β_{DTW}	1.06	2.05	1.45	1.70	1.34	0.58	0.04	-0.04	0.41	1.43	0.54	0.11	13.70	0.73	10.78	-4.31	0.46

of illiquid stocks. In other words, some small and illiquid stocks have their true beta underestimated using the standard beta measure, and under the DTW beta, the beta estimates are substantially higher. This result supports the initial hypothesis that some small, illiquid stocks may have a downward biased beta owing to sluggishness in incorporating market-wide information into their price. This result also suggests that β_{DTW} does not just estimate higher β_{DTW} for all small, illiquid stocks with low values of β, but only for those stocks where the beta should be genuinely higher. β_{DTW} still finds some small, illiquid stocks are genuine low-beta stocks.

Table 2.6 presents the equal-weighted (EW) and value-weighted (VW) excess returns associated with decile portfolios using univariate decile sorts of $\beta, \beta_{SW}, \beta_{DIM}$ and β_{DTW}. A statistically insignificant positive excess return is associated with the top-minus-bottom portfolio for β_{DTW}, whereas the three other metrics have statistically insignificant negative excess returns. Although β_{DTW} recovers portfolio returns that are increasing from low to high, the pattern is not monotonic. I find that portfolio nine has a higher average excess return than the High portfolio for β_{DTW}. Although β_{DTW} is a significant improvement over the other beta measures at recovering the CAPM predicted risk-return relation, the result is still not fully in line with the predictions of the CAPM.

I extend the study of the β_{DTW} factors from the univariate portfolio setting to the multivariate setting using Fama and MacBeth (1973) regressions. Table 2.7 presents the average coefficients from running monthly Fama and MacBeth (1973) regressions. For each measure of beta: β_{DTW}, β, β_{DIM}, and β_{SW}, four regressions are run. β in regression (ii) exhibits a statistically insignificant negative coefficient, the manifestation of the beta anomaly. β_{DIM} and β_{SW} in regressions (iii) and (iv) exhibit statistically insignificant coefficients. β_{DTW} in regression (i) exhibits a statistically insignificant positive coefficient of 0.19, reversing the negative coefficient in the β case. Each regression is then repeated when introducing a series of control

Table 2.6: Excess returns of portfolios sorted by different beta measures
Each month from July 1927 to November 2019, four sets of decile portfolios are formed by sorting stocks based on CAPM beta (β), Scholes-Williams adjusted beta (β_{SW}), Dimson adjusted beta (β_{DIM}) and DTW-adjusted beta (β_{DTW}), each calculated using rolling regressions of daily returns over the previous 12 months. This table reports the equal-weighted and value-weighted average monthly excess returns. The average monthly excess returns associated with the top-minus-bottom portfolio are reported in the final two rows. Returns are in percentages. Newey-West (1987) adjusted t-statistics using six lags are reported in parentheses for the top-minus-bottom portfolio.

	Equal-weighted				Value-weighted			
Decile	β	β_{SW}	β_{DIM}	β_{DTW}	β	β_{SW}	β_{DIM}	β_{DTW}
Bottom	0.78	0.70	0.70	0.61	0.66	0.53	0.57	0.45
2	0.84	0.84	0.75	0.74	0.63	0.61	0.60	0.56
3	0.85	0.82	0.80	0.78	0.65	0.63	0.61	0.58
4	0.89	0.87	0.85	0.83	0.66	0.63	0.67	0.72
5	0.85	0.88	0.88	0.86	0.66	0.69	0.74	0.70
6	0.90	0.90	0.90	0.88	0.72	0.77	0.77	0.73
7	0.89	0.88	0.92	0.87	0.78	0.66	0.79	0.74
8	0.80	0.85	0.92	0.87	0.70	0.76	0.75	0.69
9	0.75	0.79	0.86	0.94	0.67	0.69	0.68	0.78
Top	0.59	0.61	0.59	0.77	0.57	0.63	0.52	0.64
Top − Bottom	-0.19	-0.08	-0.11	0.16	-0.09	0.10	-0.05	0.19
	(-0.89)	(-0.56)	(-0.37)	(1.05)	(-0.40)	(-0.22)	(-0.42)	(1.10)

variables. By introducing the set of control factors, the coefficients on β, β_{DIM}, and β_{SW} are now all positive but statistically insignificant. This result is in line with Liu et al. (2018), who document that the beta anomaly arises from idiosyncratic volatility and after controlling for idiosyncratic volatility, positive loadings on beta can be obtained. The specification of β_{DTW} finds a statistically significant average slope coefficient, at the 1% confidence level when controlling for several other well-known asset pricing factors. The set of control factors does not explain the excess returns associated with the β_{DTW} factor, the interaction with these control factors strengthens β_{DTW}.

Table 2.7: Firm-level cross-sectional regressions
This table presents the average coefficient estimates from monthly Fama and MacBeth (1973) cross-sectional regressions. Each month from July 1927 to December 2019, excess stock returns are regressed on lagged predictors including DTW-adjusted beta (β_{DTW}), CAPM beta (β), Dimson adjusted beta (β_{DIM}), Scholes-Williams adjusted beta (β_{SW}) and several control variables, defined in Appendix 2.5. Each row in the table reports the time-series average of the cross-sectional regression slope coefficients and their associated Newey-West (1987) t-statistics, adjusted using six lags in parentheses. The R-squared value for each regression is reported in the far right column. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively. Statistical significance is only indicated on beta variables.

	β_{DTW}	β	β_{DIM}	β_{SW}	MAX	MIN	IVOL	ILLIQ	ISKEW	MOM	STREV	SIZE	SSKEW	TSKEW	R^2
(i)	0.19														1.65%
	(1.27)														
(ii)		-0.09													4.05%
		(-0.81)													
(ii)			0.03												3.11%
			(0.39)												
(iv)				-0.02											3.97%
				(-0.15)											
(v)					-3.91	8.57	-1.24	0.05	0.00	1.05	-0.04	-0.14	0.00	-0.05	9.01%
					(-0.95)	(2.15)	(-3.26)	(1.54)	(0.00)	(6.73)	(-9.18)	(-5.48)	(-0.26)	(-0.71)	
(vi)	0.31				-5.37	9.44	-1.14	0.05	-0.02	1.12	-0.04	-0.14	0.00	-0.02	9.40%
	(3.01)***				(-1.80)	(3.40)	(-3.55)	(2.51)	(-0.72)	(7.21)	(-9.57)	(-5.46)	(0.09)	(-0.03)	
(vii)		0.15			-5.46	10.88	-1.09	0.07	-0.04	1.05	-0.04	-0.14	0.00	0.00	10.27%
		(1.33)			(-1.34)	(2.67)	(-3.52)	(1.82)	(-0.39)	(7.55)	(-9.78)	(-5.44)	(0.07)	(-0.31)	
(viii)			0.12		-4.74	9.46	-1.19	0.05	-0.02	1.09	-0.04	-0.13	0.00	-0.02	9.85%
			(1.61)		(-1.81)	(3.13)	(-3.50)	(2.18)	(-0.60)	(7.83)	(-9.48)	(-5.60)	(-0.07)	(-0.14)	
(ix)				0.21	-5.75	10.21	-1.09	0.06	-0.03	1.07	-0.04	-0.13	0.00	-0.01	10.14%
				(1.60)	(-1.37)	(2.45)	(-3.14)	(1.59)	(-0.39)	(8.33)	(-9.03)	(-5.57)	(-0.45)	(-0.34)	

Having established that the β_{DTW} measure recovers positive excess top-minus-bottom returns, Table 2.8 presents the factor loadings and risk-adjusted alphas from regressing the returns of the top-minus-bottom portfolios for the four measures of beta on various risk models. In particular, top-minus-bottom portfolio returns for β_{DTW}, β, β_{DIM}, and β_{SW} are regressed on five different risk models: Fama–French three-factor model (FF3, Fama and French (1993)), Fama–French–Carhart four-factor model (FFC4, Carhart (1997)), Fama–French five-factor model (FF5, Fama and French (2015)), Fama–French six-factor model (FF6, Fama and French (2018)), and FFC4 with the FMAX factor (FFC4+FMAX, Bali, Cakici and Whitelaw (2011)). As the number of regressors in each risk model are increased the alphas associated with all four beta measures become less negative, and the statistical significance weakens. For β_{DTW} under the FFC4+FMAX models, statistically significant and positive alphas are obtained, whereas for the other three betas the FMAX factor renders their alphas statistically insignificant.

Table 2.8: Factor loadings and risk-adjusted alphas for top-minus-bottom beta portfolios

This table presents the factor loadings and risk-adjusted alphas from regressing top-minus-bottom portfolio returns of different beta measures against various asset pricing models. The time series of portfolio returns for beta (β), Dimson beta (β_{DIM}), Scholes-Williams beta (β_{SW}) and DTW beta (β_{DTW}) are regressed on the FF3, FFC4, FF5, FF6, and FFC4+FMAX asset pricing models. Each row in the table reports the regression slope coefficients and their associated Newey-West (1987) t-statistics, adjusted using six lags in parentheses. The R-squared value for each regression is reported in the far right column.

Factor model	Beta	α	Mkt-RF	SMB	HML	UMD	CMA	RMW	FMAX	N	R^2
FF3	β	-0.97	1.12	0.44	-0.13					1109	0.68
		(-6.12)	(21.65)	(2.38)	(-1.27)						
	β_{SW}	-0.83	0.94	0.49	0.03					1109	0.67
		(-5.85)	(18.56)	(4.18)	(0.28)						
	β_{DIM}	-0.94	1.11	0.57	0.02					1109	0.72
		(-6.22)	(21.18)	(4.30)	(0.19)						
	β_{DTW}	-0.39	0.61	0.37	0.23					1109	0.60
		(-3.18)	(12.24)	(3.53)	(2.19)						
FFC4	β	-0.81	1.09	0.43	-0.21	-0.17				1109	0.69
		(-5.04)	(23.11)	(2.38)	(-1.93)	(-2.33)					
	β_{SW}	-0.61	0.89	0.48	-0.08	-0.23				1109	0.69
		(-4.03)	(21.31)	(4.21)	(-0.81)	(-3.21)					
	β_{DIM}	-0.71	1.06	0.56	-0.09	-0.24				1109	0.74
		(-4.79)	(23.31)	(4.30)	(-0.85)	(-3.51)					
	β_{DTW}	-0.16	0.56	0.36	0.11	-0.24				1109	0.63
		(-1.25)	(14.24)	(3.46)	(1.28)	(-4.44)					
FF5	β	-0.45	0.93	0.42	-0.22		-0.75	-0.77		678	0.76
		(-2.37)	(13.16)	(5.18)	(-1.90)		(-4.23)	(-5.67)			
	β_{SW}	-0.36	0.73	0.44	-0.09		-0.69	-0.66		678	0.73
		(-2.52)	(11.32)	(6.38)	(-0.85)		(-4.39)	(-4.94)			
	β_{DIM}	-0.38	0.91	0.45	-0.19		-0.73	-0.79		678	0.77
		(-2.2)	(13.09)	(6.10)	(-1.66)		(-4.45)	(-6.07)			
	β_{DTW}	0.10	0.35	0.24	-0.07		-0.40	-0.59		678	0.58
		(0.86)	(7.52)	(4.09)	(-0.86)		(-3.21)	(-4.97)			
FF6	β	-0.33	0.90	0.43	-0.31	-0.17	-0.7	-0.72		678	0.76
		(-1.73)	(13.54)	(5.62)	(-2.97)	(-2.76)	(-4.35)	(-6.36)			
	β_{SW}	-0.20	0.69	0.46	-0.21	-0.23	-0.62	-0.61		678	0.75
		(-1.34)	(12.06)	(7.18)	(-2.66)	(-3.36)	(-4.99)	(-5.36)			
	β_{DIM}	-0.24	0.88	0.47	-0.29	-0.20	-0.66	-0.75		678	0.78
		(-1.36)	(13.65)	(6.74)	(-3.15)	(-3.03)	(-4.77)	(-6.86)			
	β_{DTW}	0.22	0.33	0.25	-0.15	-0.16	-0.35	-0.55		678	0.60
		(1.81)	(8.02)	(4.73)	(-2.40)	(-2.62)	(-3.28)	(-4.94)			
FFC4+FMAX	β	-0.09	0.65	0.12	-0.21	-0.21			0.81	678	0.84
		(-0.58)	(11.28)	(1.64)	(-3.24)	(-4.50)			(12.67)		
	β_{SW}	0.00	0.48	0.19	-0.14	-0.27			0.69	678	0.83
		(0.02)	(10.28)	(3.31)	(-2.34)	(-6.96)			(14.16)		
	β_{DIM}	-0.01	0.63	0.17	-0.19	-0.24			0.80	678	0.85
		(-0.06)	(11.75)	(2.61)	(-2.85)	(-6.21)			(14.84)		
	β_{DTW}	0.31	0.19	0.12	-0.08	-0.19			0.45	678	0.65
		(2.76)	(5.66)	(2.19)	(-1.57)	(-4.78)			(7.88)		

To further examine the characteristics of the beta measures, I perform dependent portfolio decile sorts between a set of control variables (ILLIQ, SIZE, IVOL, MOM, MAX, and BM) and the four beta measures. Table 2.9 presents the results of

these dependent portfolio sorts. Again, the FFC4+FMAX alphas for β_{DTW} under different control variables are statistically significant and positive for all controls except MOM.

Table 2.10 presents portfolio returns, FFC4 alphas, and FFC4+FMAX alphas across different size buckets and time-windows. Within each size bucket, top-minus-bottom portfolio returns for the four measures of beta are calculated. The FFC4 and FFC4+FMAX alphas of these portfolio returns are calculated and averaged across the different time windows. The main result is that in the January 1, 2000, to December 31, 2019 period, the positive relation for β_{DTW} inverts. In fact, for all beta measures across all size buckets, the portfolio returns and alphas are more negative in the January 2000 to December 2019 period compared with the July 1927 to December 1999 period. One potential explanation here is the adaptive nature of markets. As market participants learned about the beta anomaly, there could have been more market flows into low beta stocks resulting in their outperformance. In this case, correcting for asynchronicity cannot resolve the beta anomaly as it is not a result of asynchronicity between stock returns and market returns.

Table 2.9: Bivariate dependent sorts of beta measures
This table presents the results of bivariate dependent portfolio sorts between measures of beta and stock returns after controlling for selected control characteristics. Each month, all stocks are sorted into 100 portfolios, first based on the control variable and then on the selected beta measure. I present the time-series means of value-weighted excess returns of the average control variable decile portfolio within each beta decile portfolio. I also present the mean return differences of the top-minus-bottom portfolios and CAPM, FF3, FFC4, FF5, FF6, and FFC4+FMAX alphas for the top-minus-bottom portfolios. Newey-West (1987) t-statistics, adjusted using six lags, are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

Control	Beta	Low	2	3	4	5	6	7	8	9	High	H-L	CAPM α	FF3 α	FFC4 α	FF5 α	FF6 α	FFC4+FMAX α
ILLIQ	β	0.72	0.85	0.86	0.84	0.84	0.88	0.92	0.75	0.74	0.60	-0.12	-0.82	-0.85	-0.68	-0.24	-0.16	0.13
													(-5.05)***	(-6.00)***	(-4.74)***	(-1.59)	(-0.99)	(1.14)
	β_{SW}	0.67	0.80	0.87	0.86	0.93	0.87	0.80	0.87	0.76	0.58	-0.08	-0.76	-0.79	-0.57	-0.16	-0.03	0.24
													(-4.91)***	(-5.76)***	(-4.01)***	(-1.11)	(-0.21)	(2.14)**
	β_{DIM}	0.66	0.77	0.82	0.86	0.88	0.83	0.90	0.89	0.81	0.58	-0.09	-0.66	-0.69	-0.45	-0.20	-0.05	0.19
													(-4.58)***	(-5.17)***	(-3.28)***	(-1.51)	(-0.39)	(1.67)*
	β_{DTW}	0.62	0.73	0.78	0.84	0.85	0.84	0.87	0.84	0.88	0.72	0.10	-0.31	-0.37	-0.15	0.13	0.20	0.31
													(-2.68)***	(-3.19)***	(-1.28)	(1.24)	(1.73)*	(2.67)***
SIZE	β	0.73	0.81	0.85	0.90	0.89	0.92	0.88	0.86	0.75	0.58	-0.15	-0.90	-0.93	-0.76	-0.38	-0.25	0.00
													(-5.52)***	(-6.34)***	(-5.09)***	(-2.40)**	(-1.58)	(0.03)
	β_{SW}	0.62	0.83	0.86	0.87	0.94	0.92	0.89	0.87	0.76	0.60	-0.02	-0.76	-0.81	-0.58	-0.21	-0.06	0.18
													(-4.99)***	(-5.88)***	(-4.22)***	(-1.39)	(-0.39)	(1.55)
	β_{DIM}	0.65	0.76	0.83	0.85	0.91	0.93	0.90	0.89	0.84	0.59	-0.06	-0.69	-0.73	-0.51	-0.21	-0.06	0.16
													(-4.92)***	(-5.53)***	(-3.64)***	(-1.65)*	(-0.42)	(1.41)
	β_{DTW}	0.60	0.75	0.82	0.82	0.87	0.88	0.93	0.86	0.87	0.74	0.15	-0.30	-0.36	-0.16	0.12	0.21	0.30
													(-2.55)**	(-3.11)***	(-1.34)	(1.10)	(1.81)*	(2.75)***
IVOL	β	0.88	0.82	0.84	0.80	0.85	0.83	0.82	0.77	0.78	0.70	-0.18	-0.82	-0.83	-0.71	-0.40	-0.32	-0.14
													(-6.02)***	(-6.14)***	(-5.18)***	(-2.42)**	(-1.95)*	(-0.95)
	β_{SW}	0.80	0.78	0.82	0.88	0.83	0.83	0.81	0.80	0.74	0.80	0.00	-0.65	-0.67	-0.53	-0.27	-0.18	0.00
													(-5.06)***	(-5.31)***	(-4.18)***	(-1.84)*	(-1.20)	(0.04)
	β_{DIM}	0.76	0.79	0.79	0.81	0.83	0.84	0.86	0.84	0.81	0.77	0.01	-0.55	-0.58	-0.42	-0.29	-0.17	-0.02
													(-4.64)***	(-4.96)***	(-3.28)***	(-2.40)**	(-1.38)	(-0.14)
	β_{DTW}	0.71	0.73	0.75	0.80	0.81	0.85	0.85	0.82	0.89	0.88	0.18	-0.19	-0.24	-0.08	0.07	0.15	0.22
													(-1.92)*	(-2.48)**	(-0.82)	(0.73)	(1.40)	(2.02)**
MOM	β	0.84	0.88	0.91	0.86	0.90	0.91	0.84	0.80	0.75	0.51	-0.33	-0.93	-0.92	-0.91	-0.59	-0.52	-0.37
													(-6.49)***	(-6.68)***	(-6.54)***	(-3.54)***	(-3.09)***	(-2.37)**
	β_{SW}	0.75	0.87	0.86	0.87	0.87	0.89	0.89	0.85	0.75	0.58	-0.17	-0.80	-0.82	-0.76	-0.47	-0.38	-0.25
													(-6.01)***	(-6.60)***	(-6.18)***	(-3.23)***	(-2.65)***	(-1.85)*
	β_{DIM}	0.74	0.80	0.85	0.86	0.85	0.90	0.89	0.89	0.82	0.58	-0.16	-0.69	-0.71	-0.69	-0.44	-0.36	-0.24
													(-5.51)***	(-6.12)***	(-5.14)***	(-3.61)***	(-2.89)***	(-2.10)**
	β_{DTW}	0.69	0.78	0.78	0.82	0.84	0.84	0.86	0.92	0.93	0.74	0.05	-0.28	-0.33	-0.37	-0.07	-0.07	0.00
													(-2.77)***	(-3.34)***	(-3.61)***	(-0.78)	(-0.69)	(0.02)
MAX	β	0.75	0.76	0.80	0.78	0.84	0.85	0.86	0.85	0.85	0.85	0.11	-0.39	-0.36	-0.26	-0.14	-0.05	0.07
													(-2.86)***	(-2.82)***	(-1.93)*	(-0.84)	(-0.31)	(0.48)
	β_{SW}	0.69	0.78	0.76	0.83	0.77	0.80	0.88	0.85	0.90	0.93	0.24	-0.29	-0.30	-0.14	0.01	0.11	0.22
													(-2.24)**	(-2.46)**	(-1.17)	(0.04)	(0.76)	(1.65)*
	β_{DIM}	0.72	0.67	0.83	0.75	0.84	0.82	0.85	0.87	0.90	0.92	0.20	-0.24	-0.26	-0.09	-0.12	0.02	0.10
													(-2.12)**	(-2.40)**	(-0.81)	(-0.94)	(0.13)	(0.91)
	β_{DTW}	0.65	0.69	0.71	0.78	0.78	0.86	0.86	0.90	0.95	1.00	0.36	0.05	0.00	0.16	0.20	0.28	0.30
													(0.55)	(0.02)	(1.60)	(1.99)**	(2.54)**	(2.69)***
BM	β	0.71	0.72	0.75	0.76	0.75	0.76	0.77	0.74	0.71	0.66	-0.05	-0.65	-0.62	-0.46	-0.28	-0.18	0.05
													(-3.67)***	(-3.64)***	(-2.77)***	(-1.66)*	(-1.06)	(0.35)
	β_{SW}	0.63	0.71	0.73	0.80	0.76	0.72	0.82	0.74	0.70	0.70	0.07	-0.52	-0.48	-0.30	-0.13	-0.02	0.21
													(-3.04)***	(-3.04)***	(-1.96)**	(-0.88)	(-0.11)	(1.62)
	β_{DIM}	0.58	0.66	0.73	0.77	0.86	0.79	0.80	0.81	0.78	0.55	-0.03	-0.52	-0.51	-0.31	-0.24	-0.10	0.09
													(-3.37)***	(-3.74)***	(-2.26)***	(-1.77)*	(-0.72)	(0.79)
	β_{DTW}	0.51	0.64	0.75	0.77	0.78	0.79	0.78	0.81	0.85	0.64	0.14	-0.11	-0.11	0.01	0.08	0.16	0.26
													(-0.90)	(-0.96)	(0.09)	(0.71)	(1.38)	(2.25)**

Better estimates of beta can be obtained by accounting for dynamic asynchronicity between stock returns and market returns. I do not assume that the predictions of the CAPM are correct or that beta as a factor should generate consistently positive statistically significant excess returns. Instead, the DTW approach allows for a better estimate of beta where stock returns and market returns are better adjusted for than the common approaches of Dimson (1979) and Scholes and Williams (1977). There are numerous other considerations for CAPM betas in general, for example, the arbitrage of the beta anomaly with the advent of smart beta products and

investment managers actively exploiting the beta anomaly. This activity clouds the picture for the expected results for beta and whether it is a risk premium. However, using DTW to account for asynchronicity allows for the disentanglement of effects associated with beta. For example, there have been several recent studies that examine the impact of beta and risk, such as the increase of a stocks beta to an index when the stock is added to the index (Barberis et al., 2005), exchange traded funds (ETF) increasing co-movement and betas (Da and Shive, 2018), and that increased algorithmic trading increases market co-movement and betas (Malceniece et al., 2019; Park and Wang, 2020). DTW potentially allows one to disentangle whether increases in co-movement are owing to a genuine change in systematic risk or a better alignment of stock returns between the benchmark and constituents, resulting in a more accurate measure of a stock's beta.

Table 2.10: Historical performance of beta measure across size groups
This table presents the results of univariate portfolio sorts on different measures of beta split into different market capitalization and time samples. The sample is split into three size categories using NYSE breakpoints: bottom 20% (micro), next 30% (small) and top 50% (large). Within each size sample, univariate decile portfolio sorts using different measures of beta are performed. Panel A reports the average return of the difference between the high and low portfolios across different time windows. Panel B reports the FFC4 alphas. Panel C reports the FFC4+MAX alphas. Newey-West (1987) t-statistics, adjusted using six lags, are reported in parentheses. Results for the July 1927–June 1963 period are not reported in Panel C as the FMAX factor is not available.

Panel A: Portfolio returns

	All				Large				Small				Micro			
	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}
July 1927–December 2019	-0.19	-0.11	-0.08	0.16	-0.01	-0.02	0.06	0.09	-0.10	-0.01	-0.03	0.06	-0.37	-0.14	-0.07	0.37
	(-0.90)	(-0.56)	(-0.37)	(1.05)	(-0.03)	(-0.11)	(0.25)	(0.55)	(-0.40)	(-0.04)	(-0.12)	(0.34)	(-1.79)	(-0.78)	(-0.29)	(2.07)
July 1927–June 1963	0.03	0.16	0.20	0.33	0.04	0.16	0.22	0.23	0.05	0.24	0.02	0.15	-0.19	-0.04	0.20	0.79
	(0.10)	(0.51)	(0.52)	(1.14)	(0.11)	(0.43)	(0.53)	(0.70)	(0.12)	(0.75)	(0.05)	(0.54)	(-0.54)	(-0.13)	(0.49)	(2.09)
July 1927–December 1999	-0.18	-0.05	-0.05	0.26	0.31	0.19	0.26	0.18	-0.01	0.03	0.19	0.14	-0.42	0.00	-0.09	0.32
	(-0.60)	(-0.19)	(-0.18)	(1.74)	(0.93)	(0.67)	(0.80)	(1.04)	(-0.04)	(0.11)	(0.58)	(0.89)	(-1.40)	(-0.02)	(-0.30)	(2.03)
June 1963–December 1999	-0.18	-0.06	-0.05	0.25	0.32	0.18	0.27	0.18	-0.02	0.01	0.19	0.16	-0.44	-0.03	-0.11	0.29
	(-0.58)	(-0.24)	(-0.18)	(1.69)	(0.93	(0.65)	(0.82)	(1.06)	(-0.06)	(0.05)	(0.57)	(0.97)	(-1.47)	(-0.14)	(-0.36)	(1.87)
January 2000–December 2019	-0.61	-0.69	-0.65	-0.32	-0.69	-0.74	-0.59	-0.30	-0.50	-0.53	-0.52	-0.28	-0.63	-0.56	-0.49	-0.28
	(-1.16)	(-1.42)	(-1.20)	(-0.80)	(-1.16)	(-1.36)	(-1.01)	(-0.75)	(-0.83)	(-0.92)	(-0.86)	(-0.56)	(-1.29)	(-1.30)	(-0.99)	(-0.79)

Panel B: FFC4 alphas

	All				Large				Small				Micro			
	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}
July 1927–December 2019	-0.81	-0.61	-0.71	-0.16	-0.65	-0.55	-0.57	-0.30	-0.71	-0.49	-0.63	-0.20	-0.95	-0.50	-0.63	0.15
	(-5.04)	(-4.03)	(-4.79)	(-1.25)	(-3.72)	(-3.24)	(-3.39)	(-2.18)	(-3.83)	(-2.92)	(-3.69)	(-1.27)	(-4.99)	(-2.94)	(-3.93)	(0.94)
July 1927–June 1963	-0.90	-0.65	-0.73	-0.16	-0.94	-0.66	-0.74	-0.31	-0.92	-0.54	-0.97	-0.27	-1.04	-0.63	-0.57	0.48
	(-3.67)	(-2.44)	(-3.64)	(-0.89)	(-4.05)	(-2.62)	(-3.24)	(-1.69)	(-3.44)	(-2.09)	(-4.14)	(-1.23)	(-2.43)	(-1.65)	(-1.91)	(1.41)
July 1927–December 1999	-0.49	-0.27	-0.35	0.15	0.02	-0.06	-0.01	-0.06	-0.33	-0.18	-0.13	0.08	-0.76	-0.21	-0.41	0.21
	(-2.07)	(-1.60)	(-1.67)	(1.13)	(0.09)	(-0.30)	(-0.04)	(-0.38)	(-1.28	(-0.92)	(-0.56)	(0.47)	(-3.49	(-1.20)	(-1.97)	(1.40)
June 1963–December 1999	-0.49	-0.28	-0.35	0.14	0.02	-0.07	0.00	-0.05	-0.34	-0.20	-0.13	0.09	-0.78	-0.24	-0.43	0.17
	(-2.06)	(-1.68)	(-1.68)	(1.04)	(0.09)	(-0.33)	(-0.01)	(-0.36)	(-1.31)	(-1.04)	(-0.59)	(0.56)	(-3.60)	(-1.43)	(-2.09)	(1.17)
January 2000–December 2019	-1.23	-1.18	-1.25	-0.52	-1.23	-1.22	-1.15	-0.57	-1.02	-0.98	-1.00	-0.47	-1.30	-1.04	-1.15	-0.43
	(-3.64)	(-4.30)	(-4.03)	(-1.99)	(-4.03)	(-4.10)	(-3.93)	(-2.04)	(-2.64)	(-3.04)	(-2.86)	(-1.32)	(-4.11)	(-3.70)	(-4.07)	(-1.79)

Panel C: FFC4+FMAX alphas

	All				Large				Small				Micro			
	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}	β	β_{DIM}	β_{SW}	β_{DTW}
July 1927–December 2019	-0.09	0.00	-0.01	0.31	0.54	0.39	0.54	0.31	0.28	0.28	0.40	0.31	-0.48	-0.09	-0.20	0.30
	(-0.58)	(0.02)	(-0.06)	(2.76)	(3.54)	(2.69)	(3.88)	(2.48)	(1.73)	(1.87)	(2.59)	(1.86)	(-2.99)	(-0.69)	(-1.37)	(2.42)
July 1927–June 1963	0.08	0.13	0.16	0.34	0.73	0.46	0.64	0.17	0.35	0.30	0.48	0.26	-0.32	0.10	-0.01	0.36
	(0.46)	(0.94)	(0.96)	(2.57)	(4.26)	(2.58)	(3.80)	(1.23)	(1.87)	(1.71)	(2.79)	(1.63)	(-1.74)	(0.60)	(-0.07)	(2.46)
July 1927–December 1999	0.08	0.12	0.15	0.33	0.73	0.45	0.64	0.17	0.34	0.28	0.47	0.27	-0.34	0.06	-0.04	0.32
	(0.45)	(0.83)	(0.93)	(2.46)	(4.23)	(2.52)	(3.80)	(1.24)	(1.77)	(1.55)	(2.66)	(1.71)	(-1.85)	(0.38)	(-0.20)	(2.22)
June 1963–December 1999	-0.55	-0.52	-0.52	0.08	-0.09	-0.15	0.05	0.24	0.10	-0.04	0.06	0.25	-0.91	-0.65	-0.71	0.00
	(-1.82)	(-2.27)	(-1.95)	(0.40)	(-0.37)	(-0.64)	(0.21)	(1.15)	(0.30)	(-0.16)	(0.21)	(0.80)	(-2.86)	(-2.27)	(-2.68)	(0.01)

2.4 Global markets price discovery

In perfectly integrated markets securities that share commonality would exhibit no lead–lag in response to information flows. Markets would instantaneously incorporate the relevant component of new information into the price of all relevant securities. Lo and MacKinlay (1990) demonstrate that various market frictions slow the transmission of information across markets and can induce dynamic asynchronicity in markets. An example of this is a security that is listed on multiple exchanges. Suppose the exchanges on which the security is listed are all concurrently open and a relevant piece of information is released. In that case, there will be an impounding of this information into the price occurring across all exchanges. The variation in the speed at which this happens across different exchanges is what standard price discovery models attempt to measure. Thus, price discovery models explicitly deal with asynchronicity in how information is impounded into markets. However, common price discovery models typically do not account for changes in asynchronicity within the estimation window, that can lead to errors in inference when using these models. To deal with this dynamic asynchronicity problem, DTW can measure the intraday lead–lag between two assets at each time-step within the estimation window.

The link between the U.S. and U.K. equity markets is well studied, with several efforts identifying bi-directional information transmission. Eun and Shim (1989) explore the transmission of information across global stock markets using a vector autoregression framework, demonstrating the dominant influence of the U.S. stock market on other global markets, as well as bi-directional information transmission between the U.K. and the U.S. There is an extensive literature that explores volatility spillover effects that occur between different markets. Antonakakis, Floros and Kizys (2016) explore dynamic volatility spillovers between the U.K. and U.S. futures markets and identify that the relation is bi-directional and that volatility in U.K. futures are net receivers of shocks to futures volumes. However, the literature on the U.K. and U.S. futures lead–lag relation is relatively sparse. Instead, the literature focuses on the lead–lag relation between index futures and the underlying equity market indices within separate markets. The degree of the time-varying dynamic nature of the lead–lag between the U.K. and U.S. futures can be explicitly quantified by applying DTW.

2.4.1 Data and method

I use intraday quotes of the E-mini and FTSE 100 index futures to study the bi-directional transmission of information to quantify any lead–lag structures that exist.

The E-mini futures are one of the most frequently traded and liquid instruments globally and are expected to be the dominant instrument from a price discovery perspective. The E-mini futures and FTSE 100 futures, although not perfectly correlated, share a high degree of similarity as they derive their fundamental value from two of the largest developed equity markets that both respond to global macroeconomic information flows.

The continuous futures chains are sourced from Refinitiv for the E-mini (RIC: ESc1) and the FTSE 100 index futures (RIC: FFIc1). For each day between November 13, 2001, and June 30, 2020, the mid-quote is calculated using the one-second bid and ask quotes. The E-mini volume share, the ratio of the E-mini traded volume to the combined E-mini and FTSE 100 traded volume, is calculated using the total volume traded in each one-second interval. Observations where the bid exceeds the ask are replaced with the previous valid quote. In each one-second interval the log return is calculated using the mid-quote, and observations are removed where the absolute one-second log return is greater than 25%. The daily start and end times for each time series are set as the time at which the first and last trade in either the E-mini contract or the FTSE 100 contract occurs. DTW is applied to each daily time series of E-mini and FTSE 100 one-second log returns, using a DTW window of 60 seconds across all periods. Two iterations of the DTW algorithm are run to correct for the bias in the DTW algorithm. In the first iteration, the first time series provided to the DTW algorithm is the E-mini contract, and the FTSE 100 contract is set as the first time series in the second iteration. For each day in the sample, during each one-second interval two estimates of the lead–lag are obtained. The simple average of these two estimates is used as the DTW lead–lag estimate, such that a positive value corresponds to the E-mini contract leading and a negative value indicates the FTSE 100 contract leads.

Data from Refinitiv is synchronized using UTC, acknowledging that there is an inherent time delay between Refinitiv receiving information from the various exchanges and recording the data, however, this is deemed to have minimal impact at the one-second timescale used. There are several changes in the trading hours of the FTSE 100 futures within the sample period. Before June 2, 2008, the FTSE 100 contract traded between 8 am–5:30 pm London time. Between June 2, 2008, and October 4, 2010, the FTSE 100 contract traded between 8am–9pm London time. After October 4, 2010, the FTSE 100 contract traded between 1 am–9 pm London time. On November 17, 2014, ICE transitioned the FTSE 100 futures contracts from the London International Financial Futures and Options Exchange (LIFFE); at this point, the FTSE 100 contract traded between 7 am–9 pm London time. On October 1, 2015, ICE transitioned the FTSE 100 futures trading hours

to 1 am–9 pm London time. As the estimates of lead–lag are obtained for each day in the sample independently, these changing trading hours have minimal impact on the estimation procedure and only manifests when observing the pointwise lead–lag estimates.

2.4.2 Results and discussion

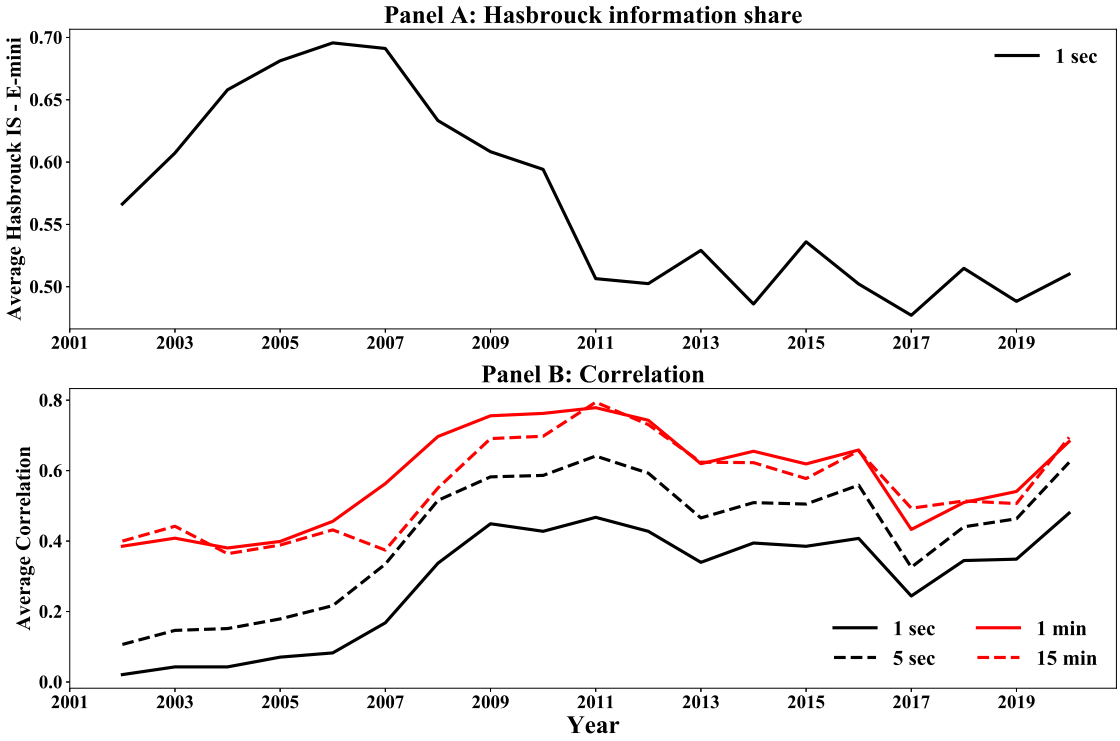

Figure 2.5: Hasbrouck information share and correlation between E-mini S&P 500 futures and FTSE 100 futures
This figure presents the Hasbrouck IS and Pearson correlation coefficients between the E-mini futures and FTSE 100 futures. Panel A presents the annual average of the daily average of the upper and lower bounds of the Hasbrouck IS. One-second mid-quote log returns are used to estimate the standard IS model each day between January 1, 2002, and May 29, 2020, between 7:00 am–5:30 pm UTC. Panel B presents the annual average of the daily correlation coefficient between the E-mini futures and FTSE 100 futures at varying frequencies of mid-quote log returns.

The standard Hasbrouck (1995) IS model is used to establish a baseline for what a standard price discovery model provides. Figure 2.5 presents the average of the upper and lower bounds estimated using the Hasbrouck IS method applied to one-second log-returns of the E-mini and FTSE 100 contracts. Before 2010, the E-mini contract was where most of the price discovery occurred, however, post-2010, the share of price discovery converges to approximately equal between both instruments. This convergence coincides with an increase in the correlation between the two instruments. The extension of the trading hours in the FTSE 100 contract that

occurred in 2009 and 2010 and a general increase in global co-movement in assets (Rua and Nunes, 2009) likely contributed to this increasing correlation. The key reason for presenting this result is to demonstrate the standard price discovery models generally only provide summary estimates of the lead–lag, often at the daily level. Although running models over finer estimation windows within a trading day is possible, there is a limit to how fine the estimation windows can be made before the model error becomes too large. One advantage of using DTW is that it is agnostic to the estimation window and can produce an estimate of the lead–lag for each observation within the estimation window.

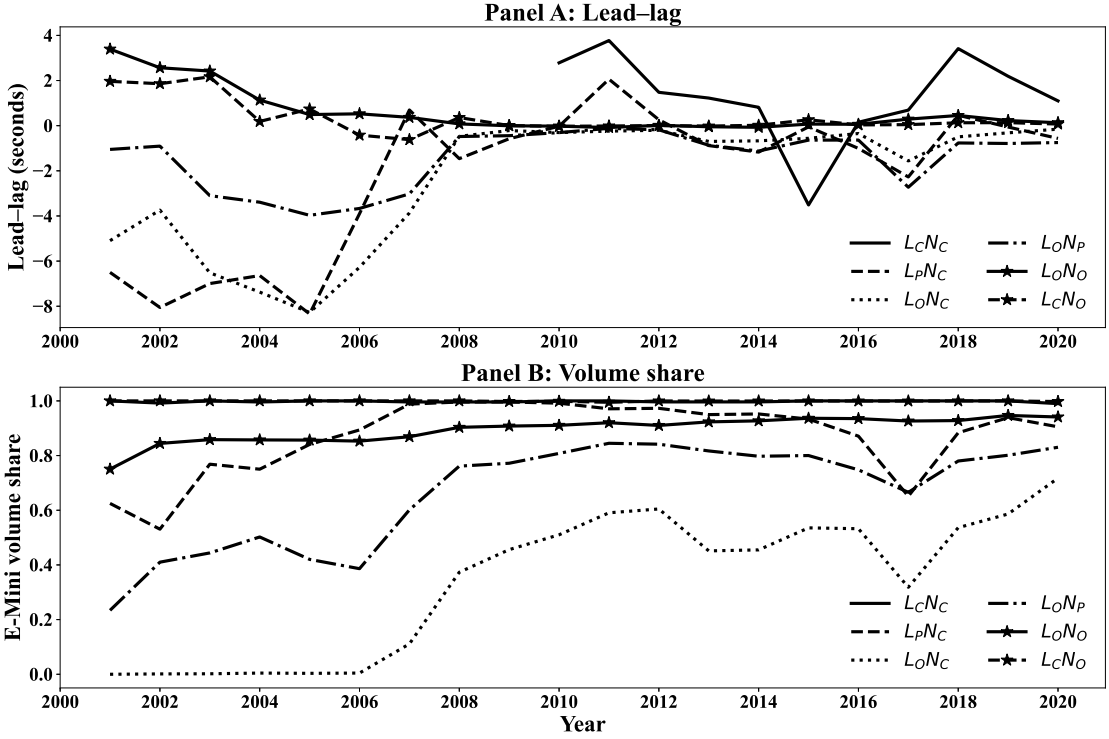

Figure 2.6: Annual DTW-estimated lead–lag between E-mini S&P 500 futures and FTSE 100 futures

This figure presents the annual average of the daily median lead–lag between the E-mini futures and FTSE 100 futures (Panel A) and the E-mini trading volume share (Panel B). For each day between November 13, 2001, and May 29, 2020, inclusive, intraday lead–lag values are estimated using DTW applied to one-second mid-quote log returns on the E-mini and FTSE 100 futures. The E-mini volume share, presented in Panel B, is the proportion of E-mini traded volume to the combined E-mini and FTSE 100 futures traded volume in each one-second interval. Each line corresponds to a combined abbreviation Xy, where X represents the underlying equity market and y represents the market operating phase. X can take values of L (LSE) and N (NYSE). y can take values of pre-open (P), open (O), and closed (C).

Figure 2.6 presents the annual average of the daily median lead–lag between the E-mini and FTSE 100 contracts and the E-mini trading volume share. Based on the operating phase of the two underlying equity markets, the trading day is divided into

six distinct periods. Each of the six periods are denoted as Xy where X can take the values L or N denoting LSE and NYSE, respectively, and y with values p, o, or c indicating pre-open, open and closed, respectively. There are several intuitive results from Figure 2.6. First, the compression of the lead–lag values toward zero over time. In particular, in the $LoNo$ period as the E-mini volume share increased from 2006, the lead that the FTSE 100 contract had over the E-mini contract reduced close to zero. This reduction toward zero is expected. With a higher level of trading volume, the prices of the two contracts will react faster to new information, and thus the lead–lag differences will be smaller. Second, the trend toward zero in the $LoNo$ period, despite the E-mini volume share remaining relatively constant. This result is potentially a function of overall increases in the volume of contracts traded and the overall increase in the speed of information transmission between London and Chicago as the infrastructure used for transmitting information has improved. Finally, in the $LcNc$ period, the E-mini lead remains above zero, indicating that the E-mini contract is leading in this period on average. The E-mini is the predominant contract used to express underlying views on overnight macroeconomic news while equity markets are closed owing to the relatively higher levels of liquidity. Although the E-mini volume share tends to be close to 100% in this period, the overall volume of contracts traded is relatively thin compared with the rest of the day.

Table 2.11: Summary statistics of intraday lead–lag between E-mini S&P 500 futures and FTSE 100 futures

This table presents summary statistics for the DTW-estimated lead–lag between E-mini futures and FTSE 100 futures. For each day from November 13, 2001, to June 30, 2020, between 1:00 am–9:00 pm UTC, I present the daily average (Mean), standard deviation (SD), skewness (Skew), kurtosis (Kurt), number of observations (Nobs), minimum (Min), maximum (Max), median ($q_{0.5}$) and 5th ($q_{0.05}$), 25th ($q_{0.25}$), 75th ($q_{0.75}$), 95th ($q_{0.95}$) percentiles. This table presents the time-series average of these daily statistics. Panel A presents summary statistics between November 13, 2001, and December 31, 2010. Panel B presents summary statistics between January 4, 2011, and June 30, 2020. Panel C presents summary statistics between November 13, 2001, and June 30, 2020. For each day in the sample, one-second mid-quote log returns on the E-mini and FTSE 100 futures contracts are used to measure the lead–lag using DTW. Each row corresponds to a combined abbreviation Xy, where X represents the underlying equity market and y represents the market operating phase. X can take values of L (LSE) and N (NYSE). y can take values of pre-open (P), open (O), and closed (C). The right-most column presents the time-series average of the average E-mini volume share, that is measured as the proportion of E-mini futures volume to the combined E-mini and FTSE 100 futures volumes, calculated using the traded volume over each one-second interval. Lead–lag values are in seconds.

Panel A: 2001–2010

	Mean	SD	Skew	Kurt	$q_{0.5}$	$q_{0.05}$	$q_{0.25}$	$q_{0.75}$	$q_{0.95}$	Volume share
LpNc	-3.95	27.28	0.02	-0.84	-3.97	-5.55	-4.76	-3.12	-2.32	0.702
LoNc	-3.15	26.48	0.09	-0.82	-4.18	-44.26	-24.22	17.93	40.06	0.296
LoNp	-2.12	26.07	0.04	-0.77	-2.15	-43.21	-22.81	18.16	39.98	0.536
LoNo	0.67	20.68	-0.01	0.23	0.89	-34.15	-13.12	14.46	34.91	0.803
LcNo	-0.23	23.06	-0.05	-0.29	0.49	-38.03	-17.17	16.46	36.70	0.998

Panel B: 2011–2020

	Mean	SD	Skew	Kurt	$q_{0.5}$	$q_{0.05}$	$q_{0.25}$	$q_{0.75}$	$q_{0.95}$	Volume share
LcNc	2.28	23.98	-0.09	0.01	1.19	-37.05	-9.45	16.24	39.57	0.960
LpNc	-0.38	24.01	-0.01	-0.46	-0.32	-39.00	-17.66	16.81	37.89	0.727
LoNc	-1.23	22.61	0.02	-0.26	-0.55	-38.80	-17.14	14.13	36.68	0.511
LoNp	-1.46	23.00	0.02	-0.33	-0.87	-39.35	-17.90	14.36	36.90	0.666
LoNo	0.28	17.92	0.04	1.06	0.10	-30.29	-10.18	10.83	31.06	0.856
LcNo	-0.30	19.48	-0.03	0.67	0.07	-34.00	-11.68	10.82	32.97	0.994

Panel C: 2001–2020

	Mean	SD	Skew	Kurt	$q_{0.5}$	$q_{0.05}$	$q_{0.25}$	$q_{0.75}$	$q_{0.95}$	Volume share
LcNc	2.28	23.98	-0.09	0.01	1.19	-37.05	-9.45	16.24	39.57	0.970
LpNc	-2.16	24.13	-0.01	-0.47	-2.14	-22.34	-11.23	6.88	17.86	0.715
LoNc	-2.19	24.54	0.05	-0.54	-2.36	-41.52	-20.67	16.02	38.37	0.405
LoNp	-1.79	24.53	0.03	-0.55	-1.51	-41.28	-20.35	16.26	38.44	0.602
LoNo	0.47	19.30	0.01	0.64	0.49	-32.22	-11.65	12.65	32.98	0.830
LcNo	-0.27	21.26	-0.04	0.19	0.28	-36.00	-14.41	13.62	34.82	0.995

Table 2.11 presents summary statistics on the lead–lag estimates, divided into the periods of 2001–2010, 2011–2020, and 2001–2020. There are no entries in Panel A for *LcNc*, as before 2009 FTSE 100 futures contract did not commence trading

until the LSE had opened at 8 am London time. First, the periods of $LcNc$ exhibit a higher standard deviation, particularly when compared with the $LoNo$ period. The estimates of DTW show that when comparing the first and second half of the sample, there has been a decrease in the standard deviation of the estimated lead–lag and a reduction in the magnitude of the lead–lag. These results demonstrate the dynamic nature of the lead–lag structures between the E-mini and FTSE 100 futures contracts. As the E-mini contract has become more heavily traded, even when the underlying NYSE is closed, it has become even more dominant in the price leadership process over the FTSE 100 futures.

One challenging result of Table 2.11 is during the $LcNo$ period the average lead–lag suggests that the FTSE 100 contract is leading the E-mini contract, whereas the median lead–lag suggests the inverse. Across this period, the lead–lag tends to be around zero except when the LSE is closing. When the LSE closes, there is typically a strong impulse response of the lead–lag in which the FTSE 100 contract takes over, followed shortly by a reversion to zero in the lead–lag. This impulse like behavior is why the mean lead–lag value suggests that FTSE 100 leads, whereas the median value suggests the E-mini leads. So far, we have only observed aggregated lead–lag statistics across different market operating phases of the LSE and NYSE. One of the strengths of the DTW technique is that it provides an estimate of the lead–lag for every observation in the estimation window. The intraday lead–lag dynamics can be observed across more granular frequencies than standard price discovery models using the DTW-provided pointwise estimates.

Figure 2.7 presents the cross-sectional average of the intraday lead–lag between the E-mini and FTSE 100 contracts, measured at a one-second frequency, in two distinct periods: 2014–2019 and 2020, using the months January through May in each year. January through May are selected to focus on the 2020 COVID-19 market events. We observe a marked shift in lead–lag behavior in 2020 owing to the COVID-19 pandemic, and the elevated levels of both market volatility and market activity. The 2014–2019 lead–lag (solid black line) shows several interesting features. The most persistent characteristic is the impulse-like responses of the lead–lag to significant intraday events, such as the opening or closing of the underlying equity markets and changes in relative levels of liquidity that occur at periods, such as the US economic announcements that occur at 1:30 pm UTC. These impulse-like responses are likely a result of the sudden sharp shifts in the trading of the futures contracts as the volumes in the underlying equity markets shift.

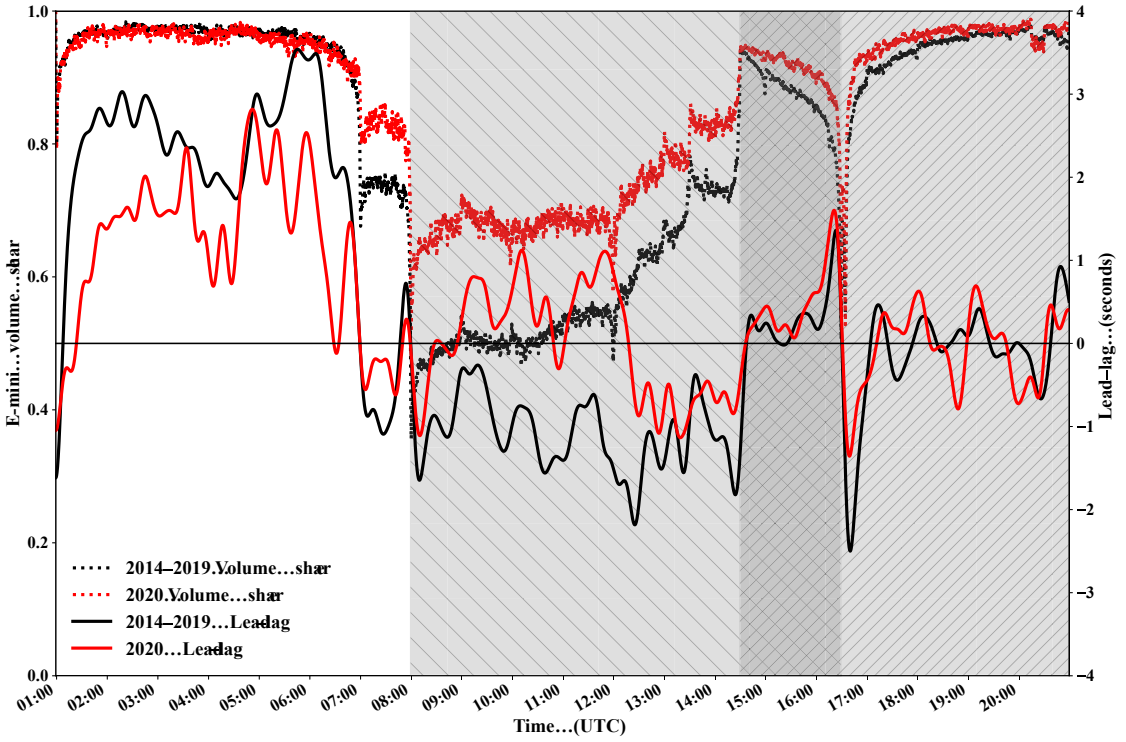

Figure 2.7: Intraday lead–lag between E-mini S&P 500 futures and FTSE 100 futures - example A

This figure presents the cross-sectional average of the DTW-estimated intraday one-second lead–lag between the E-mini futures and FTSE 100 futures (solid lines) and the cross-sectional average of the one-second intraday E-mini volume share (dotted lines). The black lines use the period of January 1, 2014, through May 31, 2019, using only the months January through May in each year. The red lines refer to January 1, 2020, through May 29, 2020. The shaded periods refer to the time at which either the LSE or NYSE are open. The left diagonal hatch is when the LSE is open. The right diagonal hatch is when the NYSE is open. Gaussian kernel smoothing is used to increase the readability of the figure.

Table 2.12 presents a qualitative description of the intraday insights that the DTW results allow, focusing on the 2014–2019 period.

Table 2.12: Description of intraday trading sessions on the NYSE and LSE

Time (UTC)	LSE	NYSE	Label	Description
1:00 am–7:00 am	Closed	Closed	*LcNc*	E-mini leads, however trading volumes in both contracts are relatively thin and there are higher levels of cross-sectional volatility in the estimates.
7:00 am–8:00 am	Pre-open	Closed	*LpNc*	The pre-open phase of the LSE sees volumes in the FTSE 100 contracts increase relative to the E-minis, and this causes the FTSE 100 to lead. Heading into the LSE open, the E-mini contract takes over before the FTSE 100 takes over just before LSE open.
8:00 am–12:00 pm	Open	Closed	*LoNc*	FTSE 100 leads across this entire period. The lead tends to be stable.
12:00 pm–2:30 pm	Open	Pre-open	*LoNp*	FTSE 100 continues to lead; however a shift occurs at 12 pm when the NYSE pre-open begins and at 1:30 pm when US macroeconomic news is announced. The E-mini contract begins to take-over heading into the NYSE open.
2:30 pm–4:30 pm	Open	Open	*LoNo*	E-mini leads, however as both markets are open, the lead tends to be stable and close to zero. Heading into the LSE close at 4:30 pm, the E-mini contract starts to take a stronger lead, before the FTSE 100 takes over.
4:30 pm–9:00 pm	Closed	Open	*LcNo*	Immediately after the LSE close, there is an impulse response where the FTSE 100 contract takes a strong lead that reverts back to zero over a 15-minute period. This is likely owing to an impounding of information from the market close, resulting in increased trading volumes and greater asynchronicity in the two instruments. For the rest of the period, the E-mini tends to lead however the overall lead–lag is approximately zero with variable noise around the estimate.

Table 2.12 focuses on the 2014–2019 period (using January through May months). I contrast this result against the 2020 lead–lag (solid red line) in Figure 2.7. During 2020, the E-mini contract tends to be more dominant when the NYSE is closed, compared with the 2014–2019 period. In particular, the E-mini volume share is elevated between 7 am–2:30 pm UTC. Across the 8 am–2:30 pm period, the difference in the E-mini volume share in the 2020 and 2014–2019 periods compress, which is reflected in the lead–lag estimates as well. The overall differential between the 2014–2019 and 2020 lead–lag is smaller during the 8 am–2:30 pm window. After 2:30 pm, the behavior of the lead–lag is quite similar in the two samples.

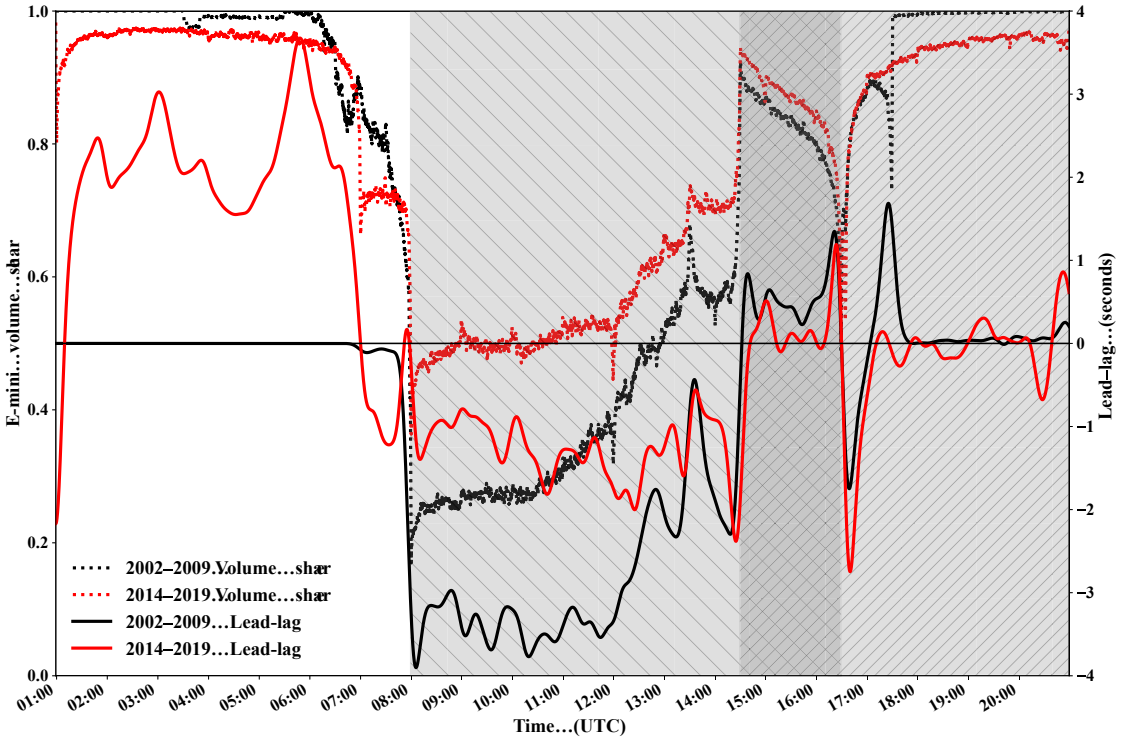

Figure 2.8: Intraday lead–lag between E-mini S&P 500 futures and FTSE 100 futures - example B

This figure presents the cross-sectional average of the DTW-estimated intraday one-second lead–lag between the E-mini futures and FTSE 100 futures (solid lines) and the cross-sectional average of the one-second intraday E-mini volume share (dotted lines). The black lines refer to the period from January 1, 2002, through December 31, 2009. The red lines refer to the period of January 1, 2014, through December 31, 2019. The shaded periods refer to the time at which the LSE or NYSE are open. The left diagonal hatch is when the LSE is open. The right diagonal hatch is when the NYSE is open. Gaussian kernel smoothing is used to increase the readability of the figure.

Figure 2.8 depicts the 2002–2009 period contrasted against the 2014–2019 period. Similar to Figure 2.7, there are several distinct periods of lead–lag behavior. In the 2002–2009 period, the futures contracts were typically only traded between 6 am–5:30 pm UTC. Between the two time periods, the overall dynamics of the lead–lag is similar, however, the magnitude of the lead–lag varies. In the 8 am–2:30 pm UTC phase during the 2002–2009 period, the lead–lag starts to move closer to zero after 12pm UTC (when the NYSE pre-open phase begins), and there are changes in the lead–lag around the 1:30pm UTC US macroeconomic news announcements. Overall, the lead–lag in the 2002–2009 period tends to be larger than that of the 2014–2019 period. This result is expected, however the intraday variation in these differentials is not constant and is worth further investigation. In particular, in the 8 am–2:30 pm period, the difference in the lead–lag between 2002–2009 and 2014–2019 is substantially larger than in the 2:30 pm–4:30 pm UTC period. This difference is likely a function of both equities markets being fully open for trading,

and thus information is being processed faster in both contracts. Overall, these examples demonstrate one of the key advantages of the DTW method. Complex intraday price leadership dynamics can be observed by applying DTW to granular windows across trading days. This technique gives a richer insight into how prices are formed in two key futures contracts, affirming the importance of the market operating phases of the equity markets underlying the futures contracts.

2.5 Conclusion

The presence of dynamic asynchronicity is an important consideration when modeling financial time series. Asynchronicity can occur across all measured frequencies of time series in financial settings, from quarterly lead–lag of macroeconomic information and asset returns to millisecond lead–lag between equity futures traded in Chicago and ETFs traded in New York. I establish DTW as a method that can quantify asynchronicity between financial time series and adjust for the time-varying nature of asynchronicity in empirical models.

Using bootstrapped simulations, I validate the use of DTW for measuring asynchronicity in financial time series. DTW proves to be robust at capturing different forms of induced lead–lag between financial time series, and I uncover insights into the influence of fundamental volatility and noise on DTW's ability to recover lead–lag structures. I explore two empirical settings at different time frequencies to demonstrate DTW's usefulness. First, I use DTW to provide a better measure of beta that accounts for dynamic asynchronicity between stock returns and market returns, helping resolve the beta anomaly. The DTW approach to beta estimation successfully recovers a positive relation between high beta stocks and excess returns. Second, using DTW, I measure the intraday dynamics in lead–lag effects between U.S. E-mini futures and U.K. FTSE 100 futures and uncover the time-varying behavior in the price leadership as the operating phase of the underlying U.S. and U.K. equity markets change.

Through these empirical studies, I demonstrate situations where DTW generates new and interesting insights into traditional financial economics problems. DTW has scope for application to other important problems where dynamic asynchronicity in financial time series can drive errors in inference in traditional empirical models.

Appendix 2.1. Variable definitions

Beta estimation

- CAPM beta (β): I follow Fama and MacBeth (1973) and measure beta using the below regression model:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i(r_{m,t} - r_{f,t}) + \epsilon_{i,t}, \tag{2.8}$$

  where $r_{i,t}$ is the return of stock i during period t, $r_{f,t}$ is the risk-free rate during period t, $r_{m,t}$ is the return of the market portfolio, and $\epsilon_{i,t}$ is the residual of stock i. The regression model specified in Eq. (2.8) is estimated using a rolling window approach. At the end of each month, the model is re-estimated using the most recent 12 months of daily returns data. The regression is only estimated if there are at least 200 daily returns observations in the most recent 12 months, and this condition applies to all measures of beta using daily data, described below.

- Dimson beta (β_{DIM}): I follow Dimson (1979) and estimate a beta that adjusts for infrequent trading events. I add five days of lagged market returns and five days of forward market returns into the regression for beta as follows:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^{(k)} \left(\sum_{k=-5}^{5} r_{m,t-k} - r_{f,t-k} \right) + \epsilon_{i,t}. \tag{2.9}$$

  The Dimson beta estimator is then the sum of the estimated beta coefficients from Eq. (2.9):

$$\beta_{DIM_i} = \sum_{k=-5}^{5} \beta_i^{(k)}. \tag{2.10}$$

- Scholes-Williams beta (β_{SW}): I follow Scholes and Williams (1977) to calculate an adjusted beta. I use three regressions to obtain three measures of betas. The first regression uses contemporaneous market returns, as in Eq. (2.8). The second regression uses the one-day lagged market excess return as the explanatory variable (Eq. (2.11)), and the third regression uses the one-day forward market excess return as the explanatory variable (Eq. (2.12)).

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^{-}(r_{m,t-1} - r_{f,t-1}) + \epsilon_{i,t}^{-}, \tag{2.11}$$

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i^{+}(r_{m,t+1} - r_{f,t+1}) + \epsilon_{i,t}^{+}. \tag{2.12}$$

The Scholes-Williams beta estimator is given by:

$$\beta_{SW_i} = \frac{\beta_i^- + \beta_i + \beta_i^+}{1 + 2\rho}, \tag{2.13}$$

where ρ is the first order autocorrelation of the market excess return over the estimation window.

- DTW beta and DTW t-statistic (β_{DTW} and β_{DTWT}): I measure β_{DTW} for each stock i during month t as follows:

    1. Start with daily stock excess returns and daily market excess returns over the months $t-11$ through t, inclusive,

        (a) (Bootstrapping step only): Randomly permute the daily stock excess returns.

    2. Calculate the compounded cumulative returns series for both the stock and the market over the estimation period,

    3. Normalize each time series by subtracting the mean and dividing by the standard deviation over the estimation period,

    4. Run DTW using a window of 20 days to obtain the optimal alignment path between the normalized cumulative stock returns series and normalized cumulative market returns series,

    5. Use the optimal alignment path from Step 4 to align the original daily stock excess returns and daily market excess returns from Step 1. The compounded cumulative returns series is not used in the estimation of beta, they are only used to obtain the optimal alignment path and then discarded,

    6. Use these optimally aligned returns series to estimate beta as in Eq. (2.8).

For each stock i in month t, I run a bootstrapping simulation using the above procedure. There is an additional Step 1(a) that is run for each bootstrapping simulation. I run $k = 100$ simulations where in each simulation I measure the $\beta_{DTWR_{i,k}}$ from permuting the stock excess returns series:

$$\beta_{DTWi} = \beta_{DTWRi} - \frac{1}{100}\sum_{k=1}^{100} \beta_{DTWR_{i,k}}, \tag{2.14}$$

$$\beta_{DTWTi} = \frac{\frac{1}{100}\sum_{k=1}^{100}\beta_{DTWR_{I,k}}}{\sqrt{\frac{\sum_{k=1}^{100}(\beta_{DTWR_{I,k}} - AVG(\beta_{DTWR_{i,k}}))^2}{100}}}. \tag{2.15}$$

This bootstrapping can also be represented in standard beta notation as:

$$\begin{aligned}
\beta_i &= \rho_{i,m}\frac{\sigma_i}{\sigma_m}, \\
\beta_i^* &= \rho_{i,m}^*\frac{\sigma_i}{\sigma_m}, \\
\beta_i^* &= \frac{1}{M}\sum_{k=1}^{K}\rho_{i,m}^*\frac{\sigma_i}{\sigma_m},
\end{aligned}$$

where β_i being the standard beta, and β_i^* being the beta obtained from permuting the stock excess returns (β_{DTWR_i}). It follows that,

$$\beta_i - \bar{\beta}_i^* = \left(\rho_{i,m} - \bar{\rho}_{i,m}^*\right)\frac{\sigma_i}{\sigma_m},$$

and theoretically, $\bar{\rho}_{i,m}^*$ should be zero.

Control variables

- Book-to-market (BM): I follow Fama and French (1992) and calculate the book-to-market ratio in month t using the firm's market value of equity as at the end of December in the previous year and the book value of common equity plus balance-sheet deferred taxes minus preferred stock for the firm's latest fiscal year ending in the prior calendar year.

- Idiosyncratic volatility (IVOL): I estimate idiosyncratic volatility using the FF3 model:

  $$r_{i,t} - r_{f,t} = \alpha_i + \beta_{MKT,i}(r_{m,t} - r_{f,t}) + \beta_{SMB,i}(SMB_t) + \beta_{HML,i}(HML_t) + \epsilon_{i,t}, \tag{2.16}$$

  where SMB_t is the return to the small-minus-big size factor portfolio, HML_t is the return to the high-minus-low value factor portfolio, and $\epsilon_{i,t}$ is the idiosyncratic return of stock i on day t. The idiosyncratic volatility of stock i in month t is defined as the standard deviation of the daily residuals using return data from the 12-month period from months $t-11$ through t, inclusive.

- Illiquidity (ILLIQ): I follow Amihud (2002) and measure illiquidity for each stock in month t as the ratio of the absolute monthly stock return to its dollar trading volume:

  $$ILLIQ_{i,t} = \frac{|r_{i,t}|}{VOLD_{i,t}}, \tag{2.17}$$

where $VOLD_{i,t}$ is the dollar trading volume of stock i on day t. Data from the 12-month period from months $t-11$ through t, inclusive, is used to calculate $ILLIQ_{i,t}$.

- Maximum return (MAX): I follow Bali et al. (2011) and measure MAX as the maximum daily return within a month.
- Minimum return (MIN): I follow Bali et al. (2011) and measure MIN as the minimum daily return within a month.
- Momentum (MOM): I follow Jegadeesh and Titman (1993) and measure momentum for each stock in month t at the cumulative return on the stock over the previous 11 months starting two months ago.
- Short-term reversal (REV): I follow Jegadeesh (1990) and Lehmann (1990) and measure short-term reversal for each stock as the return on the stock over the previous month $t-1$.
- Size (SIZE): I follow existing literature and measure firm size as the natural logarithm of the market value of equity (price times shares outstanding in millions of dollars) at the end of month $t-1$ for each stock.
- Systematic and idiosyncratic skewness (SSKEW & ISKEW): I follow Harvey and Siddique (2000) and decompose skewness into idiosyncratic and systematic components by running the following regression for each stock:

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i(r_{m,t} - r_{f,t}) + \gamma_i(r_{m,t} - r_{f,t})^2 + \epsilon_{i,t}. \tag{2.18}$$

  From this regression, ISKEW of stock i in month t is defined as the skewness of daily residuals $\epsilon_{i,t}$ from the prior 12 months. SSKEW of stock i in month t is the estimated slope coefficient $\gamma_{i,t}$.
- Total skewness (TSKEW): I follow Bali et al. (2011) and compute the total skewness of stock i for month t over the previous year t:

$$TSKEW_{i,t} = \frac{1}{D_t}\sum_{d=1}^{D_t}\left(\frac{r_{i,d} - \mu_i}{\sigma_i}\right), \tag{2.19}$$

  where D_t is the number of trading days in year t, $r_{i,d}$ is the return on stock i on day d, μ_i is the mean of returns of stock i in year t, and σ_i is the standard deviation of returns of stock i in year t.

Appendix 2.2. Additional results

Table 2.13: Adjusting for bias in the DTW algorithm
This table presents the results for a constant lead–lag simulation, in which the order of the two time series is alternated in the calculation of the lead–lag using DTW. I simulate two assets that share a common fundamental value, using the parameters, $N = 10,000$, $u_t = 1$, $s_{1,t} = s_{2,t} = 0.5$, and $W = 60$. A set of constant lead–lags are induced between the two time series in the range of -5 to $+5$ in increments of one unit. The second column contains the DTW lead–lag in which the order of the time series in the DTW algorithm is time series one first and time series two second. The third column contains the DTW lead–lag in which the order of the time series is time series two first, and time series one second. The fourth column presents the difference between column two and column three divided by two.

True lead–lag	p_1/p_2 DTW lead–lag	p_2/p_1 DTW lead–lag	Bias-adjusted lead–lag
5	4.903	-5.095	4.999
4	3.903	-4.095	3.999
3	2.904	-3.095	2.999
2	1.905	-2.096	2.000
1	0.903	-1.094	0.999
0	-0.094	-0.097	0.002
-1	-1.095	0.904	-0.999
-2	-2.097	1.906	-2.001
-3	-3.097	2.906	-3.001
-4	-4.096	3.905	-4.000
-5	-5.096	4.904	-5.000

Table 2.14: Misestimation of betas

This table presents the difference in beta estimations between the standard beta measure and Dimson, Scholes-Williams and DTW beta estimations. Each month the stock universe is divided into small, mid large stocks based on 30% lower and 70% upper cutoff values using the market capitalization of stocks on the NYSE. The value-weighted percentage differences between three beta estimation methods and the standard beta within each portfolio are calculated. I first create a Full Premia factor, that takes the difference in return between the BottomTop+TopTop portfolios and the TopBottom+BottomBottom portfolios. I then create a return set for the small universe taking the difference between the BottomTop and the TopTop portfolios. This process is repeated for large stocks and DTW beta, standard beta, Dimson beta and Scholes-Williams beta. The monthly difference in returns between the DTW factors and the other beta factors is then computed. Each row in the table reports the time-series average of the difference in return between the created risk factors and their associated Newey-West (1987) t-statistics, adjusting using six lags in parentheses. The R-squared value for each regression is reported in the far right column. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively. Values are in percent deviation from the standard beta value.

1927–1975	β_{DIM}	β_{SW}	β_{DTW}
Small	0.329	-0.078	0.739
	(7.61)***	(2.87)***	(9.25)***
Mid	0.245	-0.109	0.324
	(2.71)***	(1.45)	(2.88)***
Large	0.050	-0.196	-0.116
	(4.32)***	(-6.90)***	(-0.29)
1975–2019			
Small	1.042	0.252	1.569
	(3.78)***	(8.37)***	(6.98)***
Mid	0.340	0.061	0.175
	(7.46)***	(5.50)***	(4.04)***
Large	-0.046	-0.130	-0.218
	(-0.14)	(-0.69)	(-6.01)***
1927–2019			
Small	0.536	0.126	0.952
	(4.88)***	(7.78)***	(8.95)***
Mid	0.264	0.000	0.278
	(4.37)***	(1.93)*	(3.89)***
Large	0.004	-0.174	-0.155
	(2.26)***	(-3.58)***	(-3.13)***

Chapter 3

The index effect is not dead, it has mutated

3.1 Introduction

When a company is added to the S&P 500 index, professional investors and mainstream media alike interpret this as a positive event for the company's stock price. Academic research has thoroughly established the existence of this phenomenon, the so-called "S&P index effect." Emerging research suggests that the S&P index effect is dead, despite the significant growth of passive investing since the discovery of the effect. Now, when a stock is added to the S&P 500, the stock's price experiences no significant change. In this chapter, I find that this claim of the death of the S&P index effect is overstated. Stocks still experience significant abnormal price responses when added to the broader S&P 1500 stock universe. Whereas internal transfers of companies between the S&P 400, S&P 500, and S&P 600 no longer produce abnormal price responses. The evolving structure of passive ownership in mid- and small-capitalization companies and the presence of informed traders in the options market who seek to profit from the S&P index effect have driven this migration in the importance of different index related events.

The degree of influence that index providers, such as S&P Dow Jones Indices and MSCI Inc., have on financial markets is underappreciated. Before the open of the U.S. equity markets on December 21, 2020, Tesla, Inc. (TSLA) was added to the S&P 500.[2] Between November 16, 2020, and December 18, 2020, TSLA's share price increased 70.3%. This addition of TSLA marked the largest-ever change to the S&P 500 index, requiring passive investment strategies to purchase approximately

[2]https://www.spglobal.com/en/research-insights/articles/tesla-added-to-the-sp-500.

US$220 billion of TSLA shares, equivalent to around one-third of TSLA's total market capitalization at the time. Index providers have the power to change their policies, such as index inclusion criteria, which can have wide-ranging impacts on financial markets. For example, in 2017, S&P announced that companies who utilize multiple share class structures would not be eligible for inclusion in specific indices, directly affecting how companies approach corporate governance decisions.[3] With the ongoing growth of passive investing, the relative importance of index change events will continue to grow, creating opportunities for potential market disruption from the decisions of a select few index providers. It is thus vital to understand the implications of how index providers, such as S&P, announce and implement changes to their indexes.

Passive investment managers typically rebalance their portfolios when the index they follow changes. However, they must decide whether to rebalance immediately upon the announcement of a future change or to wait until the change becomes effective. Both decisions involve costs to the manager. Immediately rebalancing the portfolio will increase the tracking error to the index and carries the potential risk of underperformance relative to the index's return. Whereas waiting for the change to become effective can result in purchasing shares at a higher price due to other market participants anticipating increased demand, which reduces the passive portfolio's total return. The existence of an index effect can play a significant role in this decision. When stocks are added to or deleted from an index, a predictable price response can influence how passive investment managers rebalance their portfolios. Conversely, arbitrageurs can also use this information in trading strategies by anticipating and profiting from the expected increased supply or demand of shares from passive strategies. Although the index effect in the S&P 500 has been well-documented in the literature (Harris and Gurel, 1986; Shleifer, 1986; Jain, 1987; Dhillon and Johnson, 1991; Lynch and Mendenhall, 1997; Chen et al., 2004), recent studies (Kamal et al., 2012; Kim et al., 2017; Bender et al., 2019; Bennett et al., 2020) suggest a significant decline in the S&P 500 index effect.

In a market environment where passive assets under management (AUM) have increased from less than 4% in 2005 to 14% in March 2020 (Anadu, Kruttli, McCabe and Osambela, 2020), a concurrent decline in the S&P index effect challenges the existing explanations for the existence of the effect. One of the original hypotheses for the index effect, proposed by Shleifer (1986), is that demand curves for stocks are downward sloping. This hypothesis, coupled with the mechanical buying and selling that passive funds require, could explain the index effect. However, as the amount

[3]https://press.spglobal.com/2017-07-31-S-P-Dow-Jones-Indices-Announces-Decision-on-Multi-Class-Shares-and-Voting-Rules.

of assets in passive funds that mechanically buy and sell stocks has increased, why have the abnormal price responses that this trading activity is purported to create seemingly decreased? This is the primary question that I address in this chapter. First, I explore whether and how the index effect in the three headline S&P U.S. domestic equity indexes (the S&P 500, S&P 400, and S&P 600) has changed in recent years. Following this, the cross-sectional variation of the index effect is explained using measures of relative passive ownership and informed trading activity in listed options markets.

I focus on S&P's three U.S. domestic equity indexes owing to their coverage of a range of market capitalization groups and relatively opaque index change procedures. Whereas other index providers such as MSCI Inc. and FTSE Russell employ prescriptive inclusion criteria and use predetermined rebalance calendars, S&P utilizes an index committee with total discretion over final decisions of index membership. This discretionary element of S&P index changes makes them harder to predict. This prediction difficulty thus produces a more heterogeneous event sample in which arbitrage trading ahead of announcements is likely more indicative of informed trading activity, rather than hedging or risk management activity ahead of well-known index events, such as the annual FTSE Russell U.S. index reconstitution.

I collect a comprehensive sample of equity index additions and deletions for the S&P 500 (large capitalization), S&P 400 (mid capitalization), and S&P 600 (small capitalization) from 1996 to 2019, which collectively form the S&P 1500 universe. Index addition events can originate from outside the S&P 1500 or one of the other indexes, whereas deletion events are destined for outside the S&P 1500 or one of the other indexes. For instance, if a stock is promoted from the S&P 400 to the S&P 500, this is labeled as an addition by transfer. Using an event study approach, I establish that the index effect has indeed declined for the S&P 500 when measuring the effect over the full sample of additions and deletions. Before 2008, a stock added to the S&P 500 experienced an average abnormal return of 4.17% on the first trading day after the announcement day (AD). However, after 2008, this effect declined to 1.33%. This change in price response is driven predominantly by a declining price response for internal transfers from lower capitalization indexes to higher indexes. There is a marked difference when splitting the sample into stocks added to the S&P 500 from outside the S&P 1500 universe and those added to the S&P 500 that were already in the S&P 400 or S&P 600. From 2008 onward, stocks already in the S&P 400 or S&P 600 and subsequently added to the S&P 500 experienced an average abnormal AD return close to zero. In contrast, stocks added to the S&P 500 from outside the S&P 1500 universe experienced an average abnormal AD return

of 4.09%, like the pre-2008 result. Thus, the S&P index effect is still present for additions from outside the S&P 1500.

To explore the drivers of this result, I construct measures of relative passive ownership and options-based informed trading variables to explain the cross-sectional abnormal AD return variation. Relative passive ownership measures the average difference in passive ownership between the largest 500 U.S. CRSP companies and the next largest 400 companies at the end of each year, capturing the changing passive ownership structure across different company capitalizations. Measures of informed trading activity ahead of index announcements are derived from options-based abnormal call-put implied volatility (IV) spreads and abnormal volume shares of call options. The results show a significant loading on the relative passive ownership variable, particularly for index transfers from lower capitalization indexes to higher ones. A structural shift in the relative passive ownership between different stock capitalizations has occurred. The passive ownership of the top 500 companies was relatively higher than that of the following 400 companies before 2009, which has been inverted since 2009. This inversion is associated with the observed decline in the index effect for transfers between the S&P indexes.

This research contributes to the extensive literature on abnormal price responses to changes in equity indexes. Stocks earning positive (negative) abnormal returns when added (deleted) to (from) the S&P 500 is one of the most widely accepted empirical results in both academia and mainstream media. Chen et al. (2004) document statistically significant abnormal AD returns from 1962 to 2000. Since the initial studies on the S&P 500 and Dow Jones indexes, the index effect has been documented across global indexes and across different market capitalizations. A non-exhaustive list includes the U.K. FTSE 100 (Mase, 2007; Fernandes and Mergulhão, 2016), the U.S. NASDAQ 100 (Yu, Webb and Tandon, 2014; Biktimirov and Xu, 2019), the U.S. S&P 400 (Marciniak, 2010; Becker-Blease and Paul, 2010), the U.S. S&P 600 (Docking and Dowen, 2006; Gowri Shankar and Miller, 2006), the U.S. Russell indexes (Cai and Houge, 2008), MSCI World (Chakrabarti, Huang, Jayaraman and Lee, 2005), China's CSI 300 (Chu, Goodell, Li and Zhang, 2021), Toronto's TSE 300 (Masse, Hanrahan, Kushner and Martinello, 2000), and the Korean KOSPI 200 (Yun and Kim, 2010).

Following the literature documenting the S&P 500 index effect in the mid-1980s (Harris and Gurel, 1986; Shleifer, 1986; Jain, 1987) and 1990s (Dhillon and Johnson, 1991; Lynch and Mendenhall, 1997), a growing body of literature has focused on the reduction and disappearance of the index effect. Kamal et al. (2012) document a significant reduction in abnormal return for index additions in the period after the

implementation of the Regulation Fair Disclosure (October 2000), decimalization (January 2001), and the Sarbanes-Oxley Act (October 2002). Bennett et al. (2020) find, in the period 2008 to 2017, insignificant index addition returns across an 11-day window around the AD. Vijh and Wang (2022) document a trend where, during 2016 to 2019, stocks promoted from the S&P 400 to the S&P 500 have an average announcement abnormal return of -2.31% over three days, whereas stocks demoted from the S&P 500 to the S&P 400 produce a positive abnormal return of +1.21%. They propose that this result is driven by an increase in the active institutional ownership of S&P 400 stocks in recent years. My contribution to this literature finds that in recent years, the relevance of the S&P 500 index to the "S&P index effect" has declined. Instead, it is the origination of an index addition that now matters more, thereby increasing the relative importance of the S&P 400 and S&P 600 indexes.

This chapter also relates to emerging literature that uses the informational efficiency of the options market to study the abnormal behavior of the underlying equities. Augustin and Subrahmanyam (2020) and Chen, Koutsantony, Truong and Veeraraghavan (2013) both use options-based trading variables as proxies for informed trading ahead of corporate events (M&A activity) and S&P index additions. I extend this approach by utilizing IV variables that capture different properties of the IV surface. Hollstein and Simen (2021) find significantly positive responses of delta-hedged call option positions to S&P 500 addition and deletion announcements, where they focus on the impact of index announcements on the pricing of volatility in the options market. My results show the presence of abnormal options activity ahead of important corporate events (index changes) and provide further opportunity for research into the types of options-based trading strategies index arbitrageurs could be utilizing.

I also contribute to the ongoing discussion about the impact of passive ownership on price formation and the incorporation of news into asset prices. Sushko and Turner (2018) suggest that passive investing could distort securities prices and create interdependence among securities that share common passive benchmarks. Ben-David, Franzoni and Moussawi (2018) find that passive ownership increases non-fundamental volatility, whereas Glosten, Nallareddy and Zou (2021) show that ETF activity improves short-run informational efficiency. However, there is no consensus on the overall impact of ETFs on market efficiency. Similarly, Sammon (2022) uses index announcements to demonstrate that price informativeness decreases as passive ownership increases. I contribute to this debate by examining how changes in passive ownership across market capitalizations change previously established empirical results, particularly the S&P index effect. My findings emphasize the

importance of re-evaluating accepted empirical results in the presence of changing market structures and market participant behavior.

In addition to the impact that increased passive ownership has on the behavior of participants in financial markets, there is an ongoing debate on the regulatory and corporate governance implications of indexation. Rauterberg and Verstein (2013) argue that human discretion is an essential component in index creation and maintenance, even for the most mechanically created indexes, and this human element creates the potential for manipulation and malpractice. Coates (2018) examines the influence of index providers and fund managers on the concentration of company ownership, showing how this ultimately leads to the majority control of U.S. public companies landing in the hands of about 12 people. Given this human element of indexing, companies may adjust their operating behavior to meet the policy requirements of index providers for index inclusion. The increasing number of assets tracking an index increases the desirability of being included. Thus, index provider inclusion and rebalance policies can influence corporate governance and operating behavior. My results show how market phenomena, such as the index inclusion effect, can be shaped by the policy choices of index providers, which directly impacts public companies' market performance.

The results in this chapter have implications for passive investment managers. Petajisto (2011) finds an index premium of 21–28 basis points per year for the S&P 500. These are implicit costs to index investors, and minimizing these costs while maintaining the index's objectives may increase the fund's market share by delivering higher absolute returns. Knowing the extent of the index effect's presence in the S&P indexes and how it varies across different index events can help index fund managers adjust their rebalance process. Fund managers may incur a small tracking error cost by holding potential addition candidates that are not currently in the broader S&P 1500 universe, as the increased tracking error cost can be offset by reducing the impact of the index effect on the portfolio's total returns.

Moreover, these results have implications for the rebalance policies of index providers. There is considerable variation in index rebalance policies across different providers, such as S&P, FTSE Russell, and MSCI. Some providers use a mechanical rebalance process, announcing index changes well in advance following a predefined schedule with transparent criteria. Others announce changes on no predefined schedule and use more opaque inclusion criteria. My results suggest that index providers should focus on announcing the first inclusion of a stock into their broader coverage universe, as this event is likely to be the most disruptive to the stock and have the most impact on the ownership structure of a company's stock.

Finally, these results present potential opportunities for event traders. Significant price responses for additions and deletions in the S&P indexes still exist. One could form a portfolio of possible index additions based on stocks outside the S&P 1500 and benefit from the abnormal price responses.

The rest of this chapter is structured as follows. Section 3.2 details the index event sample's construction and the event study method. Section 3.3 presents the abnormal AD returns for S&P index changes. Section 3.4 presents results from a regression framework using passive ownership and options variables to explain the index effect. Section 3.5 then concludes the chapter.

3.2 Index changes sample construction

3.2.1 Data

S&P index announcements between January 1, 1996, and December 31, 2019, are obtained from the ProQuest U.S. NewsStream database. Every event related to the S&P 500, S&P 400, or S&P 600 indexes is retained.[4] Following Chen et al. (2004), the index change sample is filtered to remove stocks that will not have a sufficiently long post-event returns series and where the index change is unlikely to generate trading activity among current holders. For each event, I record the AD, effective date (ED), reported ticker, reported company name, and type of index change. Six potential events are associated with a recorded index change, with two being redundant in this study. A pure addition is an addition from outside the S&P 1500 universe to one of the three S&P indexes. A pure deletion is a complete removal of a stock from the broader S&P 1500 universe. Two events are recorded when a stock is transferred from a lower capitalization index to a higher capitalization index: one for the deletion from the current index and one for the addition to the new index. Similarly, an event is recorded for the deletion from the higher index and an addition to the lower index when a stock is demoted from a higher capitalization index to a lower capitalization index. For index promotions, the addition event is retained, and for index demotions, the deletion event is retained. This choice ensures that the expected price response (positive for additions, negative for deletions) for these events remains consistent with the addition and deletion event categories. Figure 3.1 presents the distinction between pure events and transfer events as well as the relative market capitalization ordering of the three S&P indexes.[5]

[4]Each of these indexes is mutually exclusive. The S&P 100 is not included, as it is a subset of the S&P 500.

[5]The relative market capitalization ordering between the three indexes is not exact, as there can exist reasons for an eligible stock to not be included based on its market capitalization rank.

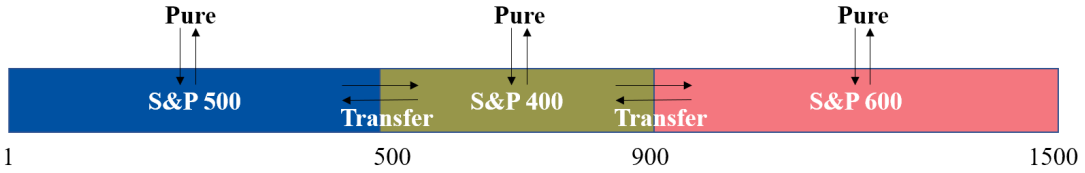

Figure 3.1: S&P index change nomenclature
This figure illustrates the hypothetical index changes that can occur across the three S&P U.S. domestic equity indexes ordered by stock rank market capitalization. Pure addition and deletion events result in a stock moving into or out of the broader S&P 1500 universe. Transfer events result in movement between the three indexes. The cutoff ranks of 1, 500, and 900 are for demonstration purposes only and can vary.

Listed U.S. options data are sourced from the OptionMetrics IvyDB U.S. database (OM) and stock data from the CRSP and Compustat databases. OptionMetrics uses their own security identifier (SECID) as the primary identifier for a company. The CUSIP identifier is used to map from the OM SECID to the CRSP PERMNO. Index announcement events are mapped to the CRSP PERMNO using the ticker and security name from the S&P index announcement. As S&P only provides a ticker and company name in their index announcements, each mapping to CRSP PERMNO is manually verified to ensure that the correct company ticker from the S&P index announcement is matched to the appropriate OM and the CRSP identifiers.

The abnormal return response of stocks to index changes is studied in two samples. The first sample, denoted as the Base Sample, is used to measure the abnormal AD returns across the different indexes. The second sample, denoted as the Regression Sample, is used to explain the cross-sectional variation in abnormal returns. To be included in the Regression Sample, a stock must have sufficient data coverage for the regression, including options data. A four-pass filter is applied to filter events for inclusion in either sample. The first pass removes events where there is no matching PERMNO in the CRSP dataset 120 days before and 30 days after the event day. The second pass eliminates events that are unlikely to generate any trading activity, such as those arising from corporate actions including mergers and acquisitions, company spin-offs, and company share class changes. These first two filters are used to construct the Base Sample. The third pass removes events where there are more than 30 days of missing options data during the model calibration window, which typically occurs between 150 and 30 days before the event. This filter is used to remove events in which a sensible estimate of expected options behavior cannot be obtained. In the fourth and final pass, events with missing data during the regression

Table 3.1: S&P index announcement sample
This table presents the frequency of index additions and deletions across the three S&P U.S. domestic equity indexes. The sample runs from January 1, 1996, through December 31, 2019. The Raw Sample is the total count of all events. The Base Sample is obtained after removing events without sufficient data in the pre- and post-event periods and events which would not require significant trading activity. The Regression Sample is all events that have sufficient market and options trading data during the event window. Pure events refer to additions to or deletions from the broader S&P 1500 universe. Transfer events refer to promotions (for additions) or demotions (for deletions) between the three indexes.

		Raw Sample		Base Sample		Regression Sample	
		Additions	Deletions	Additions	Deletions	Additions	Deletions
S&P 500	Pure	313	507	172	55	118	26
	Transfer	344	150	293	122	196	82
S&P 400	Pure	566	573	422	104	208	29
	Transfer	395	188	365	166	195	82
S&P 600	Pure	1393	1218	1273	343	394	25

window (typically from five days prior to the event day to one day after the event) are removed.

Table 3.1 shows the event count in the three samples. The Raw Sample is the count of all unfiltered index events. In total I collect 3,011 addition events and 2,636 deletion events, which constitute the Raw Sample. The Base Sample contains 2,525 additions and 790 deletion events. Finally, the Regression Sample contains 1,111 additions and 244 deletion events. Throughout the rest of this chapter, I focus on the Base Sample when analyzing aggregated abnormal return statistics.

3.2.2 Options-based informed trading variables

Empirical studies provide evidence that options trading activity reflects the variation in investor expectations about the future behavior of the underlying stock (Pan and Poteshman, 2006; Conrad, Dittmar and Ghysels, 2013; An, Ang, Bali and Cakici, 2014; Fu, Arisoy, Shackleton and Umutlu, 2016). When abnormal levels of IV variables are measured ahead of an event related to the underlying stock, it suggests that there could be informed traders using the knowledge of the event ahead of the official market announcement. These informed traders could be trying to generate higher profits by trading in the options market instead of the underlying equity market. Such trading ahead of announcements can be based on an informational advantage, such as an index prediction model, or potentially an insider trading on privileged information (e.g., SEC (2020)). The use of options variables to measure potential informed trading ahead of index change announcements aligns with these previous studies.

For each stock-day, I aggregate traded information from in-the-money (ITM), out-of-the-money (OTM), and at-the-money (ATM) option call-put pairs. Aggregate stock-day measures are constructed from each option with an open interest greater than zero, a best offer greater than zero, a bid-ask spread ratio less than 50%, an option price greater than \$0.25, and a stock price greater than \$5. All valid options are used to calculate an EW average of the IV.[6] The primary variable, the call-put IV (CPIV) spread from Fu et al. (2016), is the IV spread between ATM call and ATM put options, given their prevalence in both long bullish and long bearish strategies. In addition to IV variables, call-put traded volume is aggregated using all valid options (Chen et al., 2013).

3.2.3 Passive ownership

To measure the aggregate passive ownership, I follow Doshi, Elkamhi and Simutin (2015) and obtain fund holdings from the Thomson Reuters Financial Mutual Fund Holdings (s12) database, linking this with the CRSP database using the WRDS MFLINKS database. From the s12 database, the latest available holdings for each passive ETF are obtained from January 1, 1996, to December 31, 2018. The s12 database is merged with the CRSP Survivor-Bias-Free Mutual Fund database using MFLINKS. ETFs that have s12 investment objective codes of 1, 5, 6, or 7 are retained. These ETFs are filtered for CRSP share code (SHRCD) 73 (exchange-traded products), and domestic U.S. equity ETFs are obtained using the CRSP objective code and exchange-traded flag.

The lagged one-month market capitalization for all stocks in the U.S. CRSP universe is calculated using the CRSP shares outstanding (SHROUT) and PRC fields and used to construct an annual measure of the change in passive ETF ownership. This market capitalization is merged with the ETF holdings using the CRSP PERMNO and DATE fields. For each year-month pair, the sample is cross-sectionally ranked based on market capitalization (rank one denoting the largest stock in the universe). For each year, the average passive ETF ownership for each stock rank is calculated by averaging across the rank for each month in the given year. Finally, the passive ownership spread between the top 500 companies and next 400 companies is given by:

$$PASSOWN_t = OWNR_{t,1-500} - OWNR_{t,501-900}, \tag{3.1}$$

[6]The results are not sensitive to using an open interest weighted average.

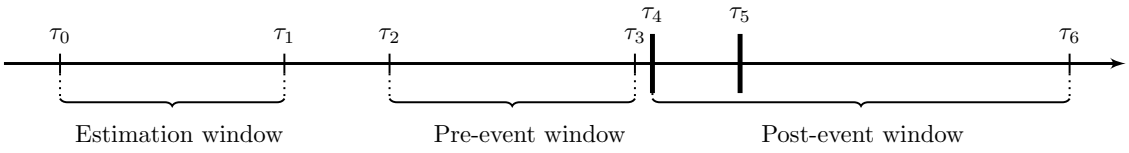

Figure 3.2: Index announcement event study timeline
This figure summarizes the time periods used in the event study approach. τ_0 is the start of the estimation window, τ_1 is the end of the estimation window, τ_2 is the start of the pre-event window, τ_3 is the end of the pre-event window, τ_4 is the event day, τ_5 is the ED of the event, and τ_6 is the end of the post-event window. The estimation window occurs prior to the event and is used to calibrate market-based models. The pre-event window occurs after the estimation window and ends the day before the event. Variables of interest during the pre-event window are used to explain the abnormal return variation on the event day (τ_4) and in the post-event (τ_4 to τ_6) period.

where $OWNR_{t,1-500}$ is the average passive ETF ownership in year t for stocks ranked between 1 and 500, and $OWNR_{t,501-900}$ is the average passive ETF ownership in year t for stocks ranked between 501 and 900.

3.2.4 Event construction and abnormal values

The primary variable of interest is the AD abnormal stock return response to S&P index changes. An event study method is used to measure abnormal values. Figure 3.2 depicts the different periods used in the event study. The first trading day after the announcement is the event day (τ_4). Expected value models needed to measure abnormal values are estimated using the period from 150 days (τ_0) to 31 days (τ_1) before the AD. In this study, two market models are used. The first is a CAPM-based market model to measure abnormal stock returns:

$$r_{i,t} = \alpha_i^s + \beta_i^s r_{m,t} + \epsilon_{i,t}, \tag{3.2}$$

where $r_{i,t}$ is the stock i return on day t, $r_{m,t}$ is the market return on day t, $t = \tau_0, \ldots, \tau_1$, and $\mathbb{E}(\epsilon_{i,t}) = 0$. The main set of results presented is based on $\tau_0 = -150$, $\tau_1 = -31$, and pairs of $(\tau_2, \tau_3) = \{(-10,-1), (0,0), (1,10), (-10,10)\}$. The AD is denoted by the pair: $(\tau_2, \tau_3) = (0,0)$. The market return corresponding to the relevant index for each event is used (i.e., $r_{m,t}$ for an S&P 400 index event is the S&P 400 index return). Using the estimates of $\hat{\alpha_i}^s$ and $\hat{\beta_i}^s$ from Eq. (3.2), the abnormal stock return is calculated as:

$$AR_{i,t} = r_{i,t} - (\hat{\alpha}_i^s + \hat{\beta}_i^s r_{m,t}), \tag{3.3}$$

where $t = \tau_2, \ldots, \tau_3$ and $\tau_2 \geq \tau_1$. A second market model is used to measure the abnormal level of IV and volume-based variables. Using the same estimation window as the previous model, the model is specified as:

$$V_{i,t} = \alpha_i^v + \beta_{i,m}^v V_{m,t} + \epsilon_{i,t}, \tag{3.4}$$

where $V_{i,t}$ is the variable of interest for stock i on day t, $V_{m,t}$ is the VW market average of the variable of interest on day t, and $t = \tau_0, \ldots, \tau_1$. The abnormal value of $V_{i,t}$ is calculated as:

$$AV_{i,t} = V_{i,t} - (\hat{\alpha}_i^v + \hat{\beta}_{i,m}^v V_{m,t}), \tag{3.5}$$

where $t = \tau_2, \ldots, \tau_3$ and $\tau_2 \geq \tau_1$.

3.2.5 Volume ratios

To accurately measure abnormal volume ratios, the market's overall volume needs to be accounted for. A single stock's trading volume could be elevated simply because a large amount of trading volume in the general market spills over into the single stock. The overall market volume level is accounted for using an incremental turnover ratio. First, the stock-level volume share is calculated as the ratio of the daily traded volume to the current shares outstanding. Next, capitalization-weighted market volume share is calculated by considering all valid stocks in the CRSP or OM sample. The turnover ratio is calculated as the ratio of the stock volume share to the capitalization-weighted market volume share. Finally, the abnormal volume share is calculated by subtracting the average of the volume share during the calibration window from the volume share during the event window.

3.3 Abnormal stock responses to index changes

3.3.1 Aggregate additions and deletions

Abnormal returns to index additions and deletions are calculated in an event study framework using the first valid trading day after the announcement of the relevant change (as S&P announces index changes after market close). Figure 3.3 presents each index's yearly average abnormal returns for additions (Panel A) and deletions (Panel B). Regarding the index additions in Panel A, there are two noteworthy patterns. First, abnormal returns for the additions to the S&P 400 and S&P 500 indexes turned marginally negative in 2018 and 2019. Second, abnormal returns

associated with additions to the S&P 600 index have increased in recent years, reaching over 8% on average in 2019. The findings for the S&P 500 and S&P 400 are consistent with recent literature highlighting the disappearance of the index effect in this these indexes (Kamal et al., 2012; Kim et al., 2017; Bender et al., 2019), with some studies even reporting a negative effect.

When an index addition occurs, there is typically a one-for-one deletion from the index. Deletions generally initiate index changes, but index deletions tend to be driven by activity that does not require trading from the shareholders. For example, a company being acquired or a bankruptcy filing that results in the suspension of the stock from the exchange it is listed on. As such, the sample size for index deletions is typically smaller than that for additions, which can lead to biases in subsequent analyses. Panel B in Figure 3.3 shows an increase in the abnormal returns toward zero for index deletions; although not an exact symmetric result to the trend in the additions, it is similar. We observe a similar trend in the S&P 600 deletions as in Panel A, where the abnormal returns associated with the S&P 600 deletions have been getting more negative in 2017–2019. Again, these results give credence to the death of the S&P index effect, as deletions from the S&P 400 and S&P 500 no longer result in negative abnormal returns. However, this result can be juxtaposed to the S&P 600, where we observe a strengthening index effect.

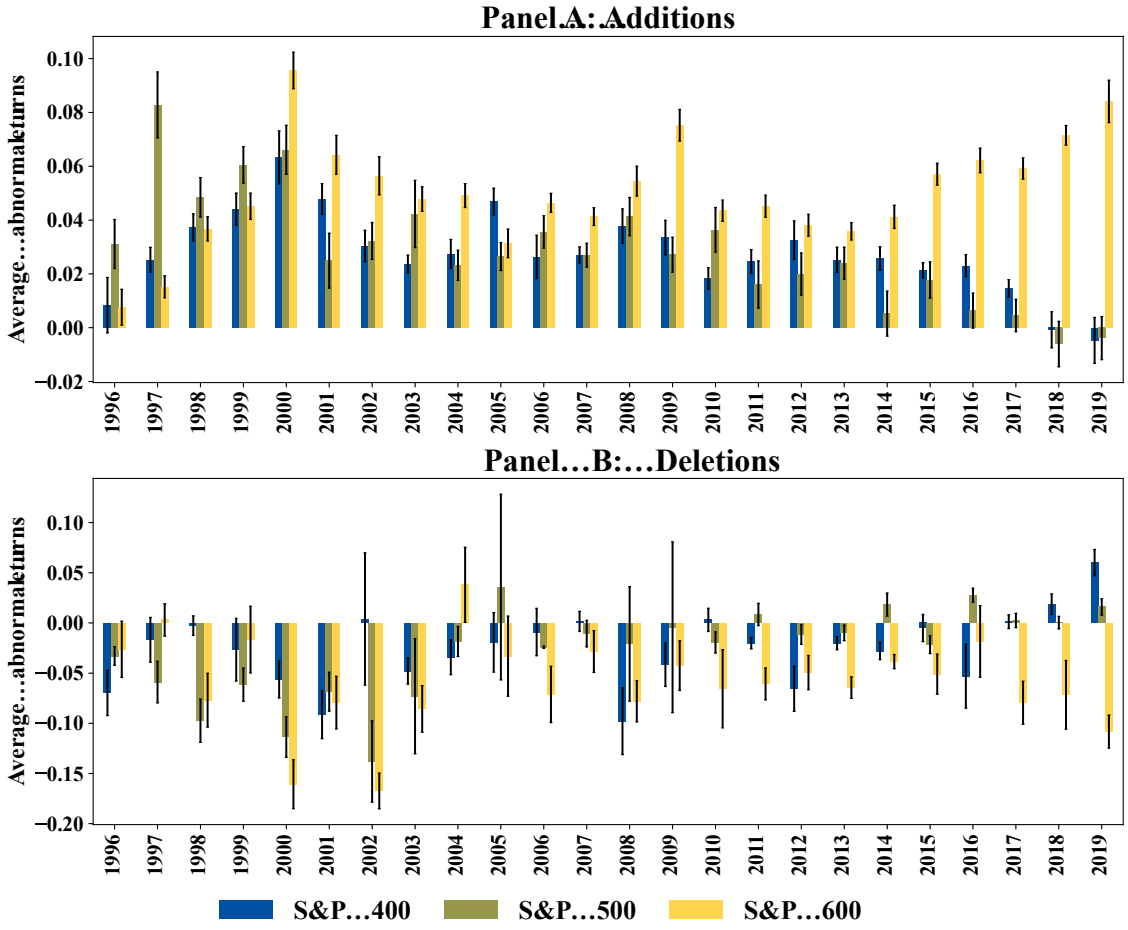

Figure 3.3: Annual average announcement day abnormal returns
This figure displays the annual cross-sectional average of the abnormal AD returns for addition and deletion announcements of S&P indexes from January 1, 1996, to December 31, 2019. The Base Sample is split into the S&P 400, S&P 500, and S&P 600 announcements. The annual average abnormal returns are plotted for the AD, and a standard error bar is included to represent the variability of the abnormal returns across the sample.

These full sample results highlight the reduced abnormal AD returns associated with additions and deletions to the S&P 500 do not show the complete picture. I extend the analysis by considering internal transfers between S&P indexes. Any movement of companies between indexes is also reported when an index announcement is made. For example, suppose a stock was already in the S&P 600 and the announcement stated that the stock was being added to the S&P 400. In that case, the announcement also records a corresponding deletion event for the S&P 600 and often explicitly states the company as an "*existing company in the S&P 400/500/600 index.*" Using this information, I construct a full sample of index promotions (stocks moving from a lower capitalization index to a higher capitalization index) and index demotions (stocks moving from a higher

capitalization index to a lower capitalization index). For additions, I only use index promotions, and for deletions, I use only index demotions. An addition to an index can arise from a demotion (e.g., a stock demoted from the S&P 500 to the S&P 400 has been deleted from S&P 500 and added to the S&P 400), and a deletion can arise from a promotion: I use the convention of additions being indicative of a positive event and deletions being indicative of a negative event. This choice prevents the duplication of events across the different indexes and ensures that each event can only enter the complete sample once.

3.3.2 Transfers between indexes

Figure 3.4 presents the results of separating the index additions into upward transfers (Panel A) and pure additions (Panel B). These two sub-samples of additions exhibit very different abnormal returns. First, pure additions in Panel B have maintained a stable average AD abnormal return of 4.70% across the three indexes from 2001, and we do not observe any significant decreases in recent years. On the other hand, upward transfers in Panel A show a significant change. Before 2011, the two indexes had a relatively low average abnormal return of 2.20%. From 2011, the abnormal return averaged -0.14%, and in 2018 and 2019, it turned negative. These differing effects can only be observed by splitting the additions sample. Although these results are reasonably clear, the relative number of events that occur in the transfer and pure addition sub-samples can influence the economic relevance of these results. If most of the index changes classified as additions are now upward transfers, then a statistically significant positive abnormal return for pure additions would have little relevance in the future.

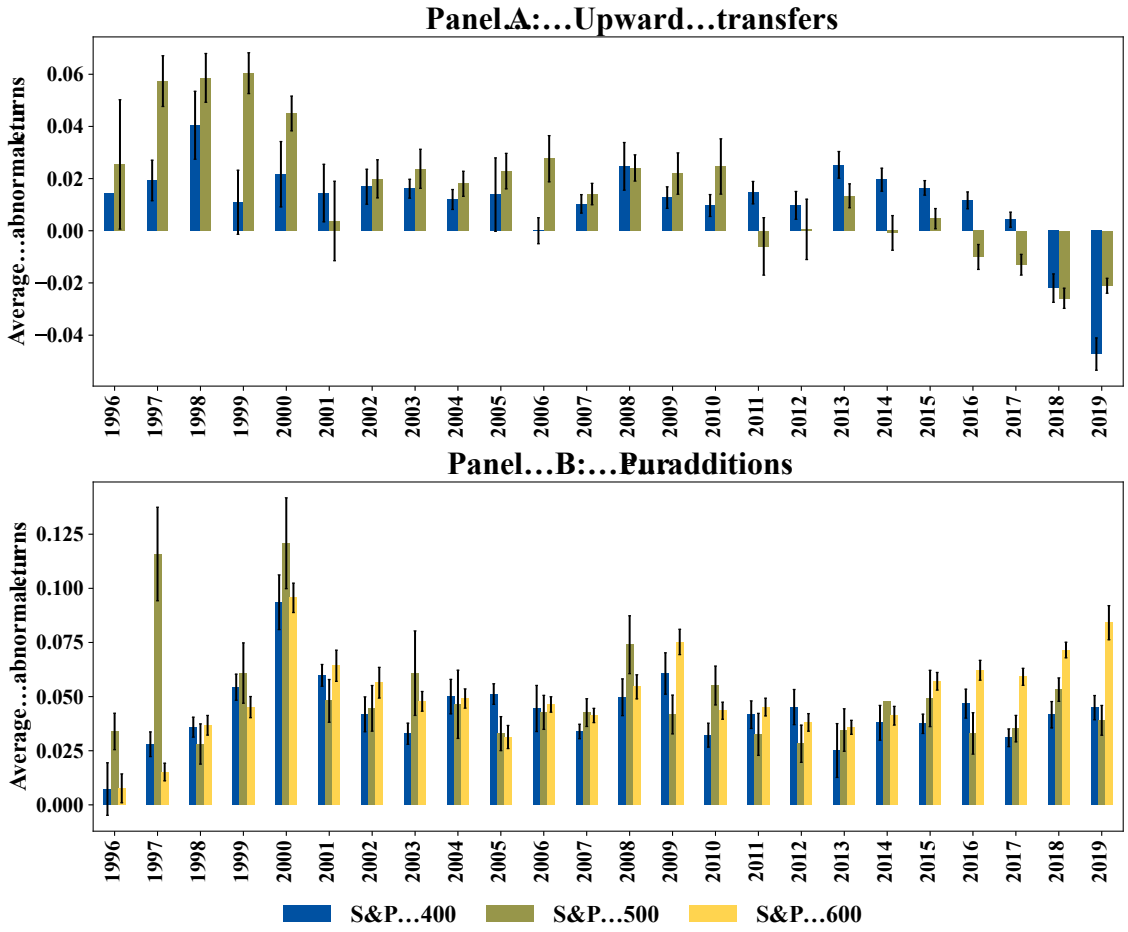

Figure 3.4: Average announcement day abnormal returns for index additions
This figure displays the yearly cross-sectional average of the abnormal AD return for S&P index addition announcements from January 1, 1996, to December 31, 201. Panel A presents the results for upward transfers (e.g., a stock in the S&P 600 is promoted to the S&P 400). Panel B presents the results for pure additions. The Base Sample is split into the S&P 400, S&P 500, and S&P 600 announcements. The annual average abnormal returns are plotted for the AD, and a standard error bar is included to represent the variability of the abnormal returns across the sample.

Figure 3.5 presents the annual frequency of index change events. In Panel A, the additions are split into two categories: pure additions or stocks added from outside the S&P 1500 universe and transfers (index promotions). Trends among the three indexes are identified by partitioning the sample by index. One clear trend is the decreasing proportion of pure additions for the S&P 400 and S&P 500 indexes. Before 2009, the average frequency of pure additions was approximately 50%; after 2009, it declined to approximately 25%. This change suggests that the S&P index committee has opted to promote more stocks through the S&P 600 and S&P 400 indexes and finally into the S&P 500 rather than directly adding companies to the flagship S&P 500 index. Further, the proportion of pure additions to the S&P 400 has only slightly declined, with a larger proportion of additions coming from

index demotions from the S&P 500 in recent years. Panel B presents the deletions, separated into full deletions and demotion transfers. The declining proportion of full deletions in the S&P 500 and S&P 400 and the increasing proportion of index transfers are clear trends. These trends again indicate that the S&P index committee prefers to sequentially demote stocks through each index instead of immediately removing them from the S&P 400 or S&P 500. Overall, there is a clear changing frequency of different index events that influences each event's economic relevance in the sample. The ongoing presence of pure additions in the latter half of the sample means that these events are still relevant, and the results in Figure 3.4 are worth further investigation.

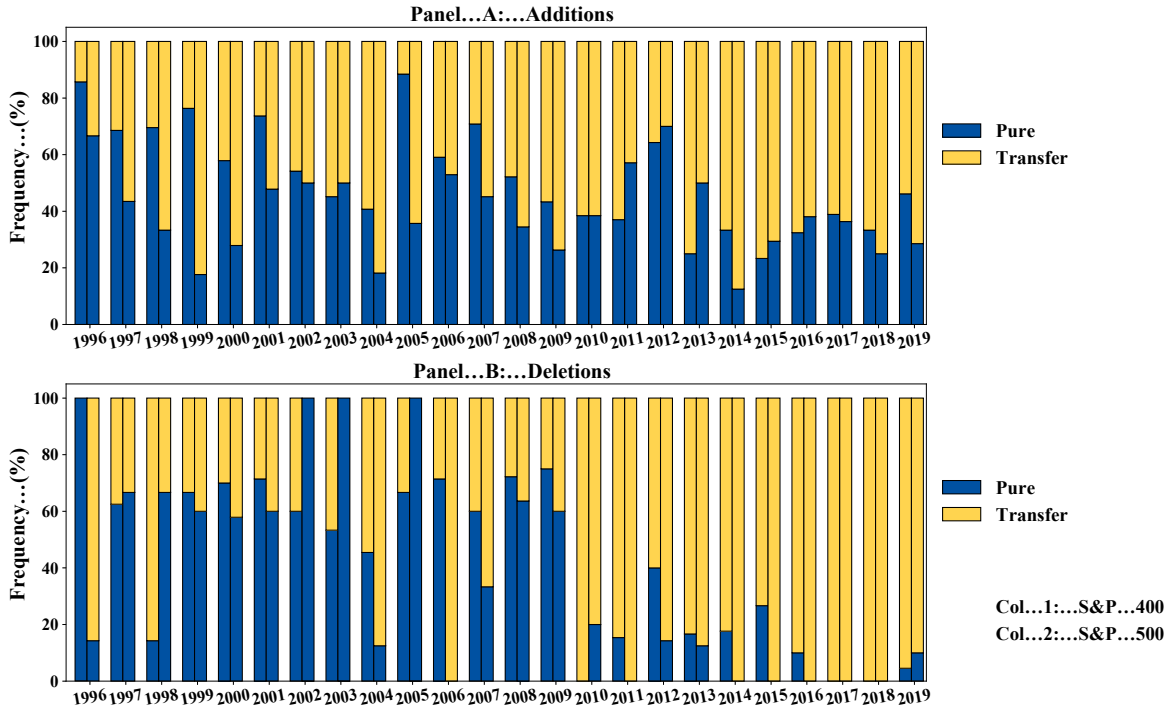

Figure 3.5: Proportional frequency of S&P index change events
This figure presents the annual proportion of S&P index change events for the Base Sample from January 1, 1996, to December 31, 2019. Panel A presents the results for index additions, where stocks are categorized into new additions to the S&P universe (pure) and stocks that are transferred between indexes (transfer). Panel B presents the results for index deletions. In both Panel A and Panel B, for each year, the first column corresponds to the S&P 400 and the second column to the S&P 500. Results for the S&P 600 are not presented, as all events are pure additions/deletions.

Table 3.2 presents the AD average abnormal returns for index additions and deletions of the three U.S. S&P indexes. Three categories are used to partition the sample: pure, transfer, and a combined full sample. Two sub-samples of time are used to partition the sample: January 1, 1996, through December 31, 2008 (1996–2008) and January 1, 2009, through December 31, 2019 (2009–2019). The results show

that although the magnitude of the abnormal return response has decreased for pure additions, it remains statistically significant. However, abnormal returns in the 2009–2019 sample are statistically insignificant for index transfers. There is no statistically significant change in sub-periods for additions to the S&P 600, where the average abnormal return increases from 5.04% to 5.71%. In addition, I find that the Vijh and Wang's (2022) result of strongly positive returns associated with additions to the S&P 500 from outside the S&P 1500 holds for the S&P 400 and S&P 600, with both exhibiting statistically significant positive abnormal returns for pure additions. These results generally hold for deletion events. The primary difference from the additions results is that the pure deletions for the S&P 500 in the 2009–2019 sub-sample are no longer statistically significant but are marginally positive. Given the relatively small sample size available for deletion events, drawing conclusions is difficult. In sum, I conclude that the index effect is alive and statistically significant for pure additions, whereas for index transfers, there are no longer any statistically significant abnormal return responses. Thus, what matters now is inclusion into the broader S&P 1500 universe. However, once a stock has been included into one of the S&P 400 or S&P 600, promotion into the S&P 500 is not as important as it once was.

Table 3.2: Announcement day abnormal returns
This table presents the average abnormal AD return of S&P index addition and deletion events. The sample runs from January 1, 1996, through December 31, 2019, and uses the Base Sample. Stocks are classified as those that enter or leave the S&P 1500 universe (pure) or stocks that internally move between indexes (transfer). The sample is divided into two sub-periods: January 1, 1996, through December 31, 2008, and January 1, 2009, through December 31, 2019. Abnormal returns are reported in percentages. t-statistics are reported in parentheses. t-statistics reported for the difference in average abnormal returns ((3) − (2)) assume unequal sample variance. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively, and only for differences in values.

		Additions			Deletions		
		S&P 500	S&P 400	S&P 600	S&P 500	S&P 400	S&P 600
Pure	1996–2019 (1)	5.27	4.83	5.29	-6.70	-5.90	-7.20
		(15.76)	(23.88)	(41.21)	(-3.01)	(-5.34)	(-11.64)
	1996–2008 (2)	6.04	5.06	5.04	-8.40	-6.09	-7.94
		(12.35)	(19.19)	(27.79)	(-4.31)	(-4.53)	(-9.03)
	2009–2019 (3)	3.98	4.20	5.72	0.91	-5.36	-5.90
		(13.40)	(19.18)	(35.84)	(0.11)	(-2.85)	(-8.33)
	(3) − (2)	-2.06***	-0.74**	0.68***	9.30	0.73	2.05*
		(-3.43)	(-2.13)	(2.82)	(1.07)	(0.32)	(1.81)
Transfer	1996–2019 (1)	2.22	1.00		-1.13	0.01	
		(9.69)	(5.31)		(-3.51)	(0.02)	
	1996–2008 (2)	3.69	1.82		-3.74	-1.07	
		(12.98)	(5.77)		(-6.84)	(-1.93)	
	2009–2019 (3)	-0.27	0.26		0.05	0.44	
		(-1.14)	(1.27)		(0.14)	(1.03)	
	(3) − (2)	-3.97***	-1.57***		3.79***	1.52**	
		(-10.33)	(-4.17)		(5.95)	(2.16)	
Combined	1996–2019 (1)	3.35	3.05	5.29	-2.86	-2.27	-7.20
		(16.59)	(19.71)	(41.21)	(-3.83)	(-4.48)	(-11.64)
	1996–2008 (2)	4.58	3.90	5.04	-6.26	-4.17	-7.94
		(17.45)	(18.09)	(27.79)	(-5.65)	(-4.73)	(-9.03)
	2009–2019 (3)	1.25	1.73	5.71	0.14	-0.64	-5.90
		(5.22)	(9.26)	(35.84)	(0.15)	(-1.20)	(-8.33)
	(3) − (2)	-3.33***	-2.18***	0.68***	6.40***	3.53***	2.05*
		(-9.37)	(-7.63)	(2.82)	(4.47)	(3.43)	(1.81)

Table 3.2 solely focuses on the AD abnormal returns. Table 3.3 presents cumulative abnormal returns for different event windows. The S&P 500 exhibits substantially different behavior from the other two indexes when looking at different accumulation windows. In the window of 10 days before and 10 days after the announcement ($\tau = [-10, 10]$), the cumulative abnormal return of pure S&P 500 additions significantly drops and is borderline statistically significant in the 2009–2019 period. This result for the S&P 500 contrasts with the S&P 400 and S&P 600, which is either at a similar level or higher in the 2009–2019 period than in the 1996–2008 period for the $\tau = [-10, 10]$ window. The returns in the 10 days following the AD drive this result for the S&P 500, where roughly one-third of the AD abnormal return reverts.

Arbitrageur shareholders selling down positions to profit from the previous price increases, resulting in a price reversion owing to an excess supply of stock in the market, could drive this behavior.

Table 3.3: Cumulative abnormal returns for different event windows
This table presents the average cumulative abnormal returns across different event windows for S&P index addition and deletion events. The sample runs from January 1, 1996, through December 31, 2019, and uses the Base Sample. Stocks are classified as those that enter or leave the S&P 1500 universe (pure) or stocks that internally move between indexes (transfer). The sample is divided into two sub-periods: January 1, 1996 through December 31, 2008, and January 1, 2009, through December 31, 2019. Abnormal returns are reported in percentages. t-statistics are reported in parentheses.

Panel A: Additions													
	Index	S&P 400				S&P 500				S&P 600			
	Window	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$
Pure	1996–2008	-1.38	5.06	-0.29	3.39	1.36	6.04	0.82	8.22	-0.56	5.04	-1.73	2.75
		(-2.56)	(19.19)	(-0.50)	(4.90)	(1.29)	(12.35)	(0.94)	(5.67)	(-1.58)	(27.79)	(-4.42)	(5.62)
	2009–2019	-0.69	4.20	-0.19	3.32	-0.20	3.98	-1.35	2.43	-0.20	5.72	-1.48	4.04
		(-1.07)	(19.18)	(-0.36)	(4.80)	(-0.35)	(13.40)	(-1.91)	(2.51)	(-0.59)	(35.84)	(-4.37)	(8.62)
Transfer	1996–2008	-0.33	1.82	-2.02	-0.53	0.57	3.69	-2.07	2.20				
		(-0.48)	(5.77)	(-2.46)	(-0.51)	(0.88)	(12.98)	(-3.32)	(2.54)				
	2009–2019	-0.61	0.26	-1.92	-2.27	-0.62	-0.27	-1.87	-2.77				
		(-1.59)	(1.27)	(-5.02)	(-3.92)	(-1.30)	(-1.14)	(-4.02)	(-3.83)				
Panel B: Deletions													
	Index	S&P 400				S&P 500				S&P 600			
	Window	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$	$[-10,-1]$	$[0,0]$	$[1,10]$	$[-10,10]$
Pure	1996–2008	-9.66	-6.09	3.39	-12.36	-19.41	-8.40	10.97	-16.83	-18.72	-7.94	11.37	-15.30
		(-3.23)	(-4.53)	(1.31)	(-3.81)	(-4.10)	(-4.31)	(1.98)	(-3.09)	(-8.69)	(-9.03)	(4.71)	(-5.41)
	2009–2019	-11.79	-5.36	6.39	-10.76	-3.37	0.91	2.06	-0.41	-7.66	-5.90	6.59	-6.97
		(-2.51)	(-2.85)	(0.77)	(-1.53)	(-0.67)	(0.11)	(0.31)	(-0.05)	(-3.66)	(-8.33)	(2.56)	(-2.34)
Transfer	1996–2008	-0.69	-1.07	0.92	-0.84	-1.44	-3.74	-0.00	-5.19				
		(-0.69)	(-1.93)	(1.12)	(-0.58)	(-0.86)	(-6.84)	(-0.00)	(-2.94)				
	2009–2019	0.52	0.44	3.00	3.96	0.30	0.05	1.19	1.54				
		(0.55)	(1.03)	(2.94)	(2.88)	(0.43)	(0.14)	(1.50)	(1.26)				

Figures 3.6 and 3.7 depict the average cumulative abnormal return across several sub-periods and samples from 30 days before ($t = -30$) and 30 days after ($t = +30$) the index announcement. Figure 3.6 uses the cumulative abnormal return when considering pure and transfer index events. Panel A of this figure focuses on additions and highlights a clear decrease in the abnormal returns at $t = 0$ (AD) for the S&P 400 and S&P 500. However, we also observe a significant reversion in abnormal returns after the event day in both samples. The relative underperformance of a stock after its inclusion in the index could explain this result. The stock may have experienced significant positive returns in the window preceding index inclusion. However, upon inclusion into the index, its behavior shifts to be more in line with the rest of the index. Consequently, this results in an underperformance of the stock relative to its calibration window beta. Da and Shive (2018) have documented this effect in detail.

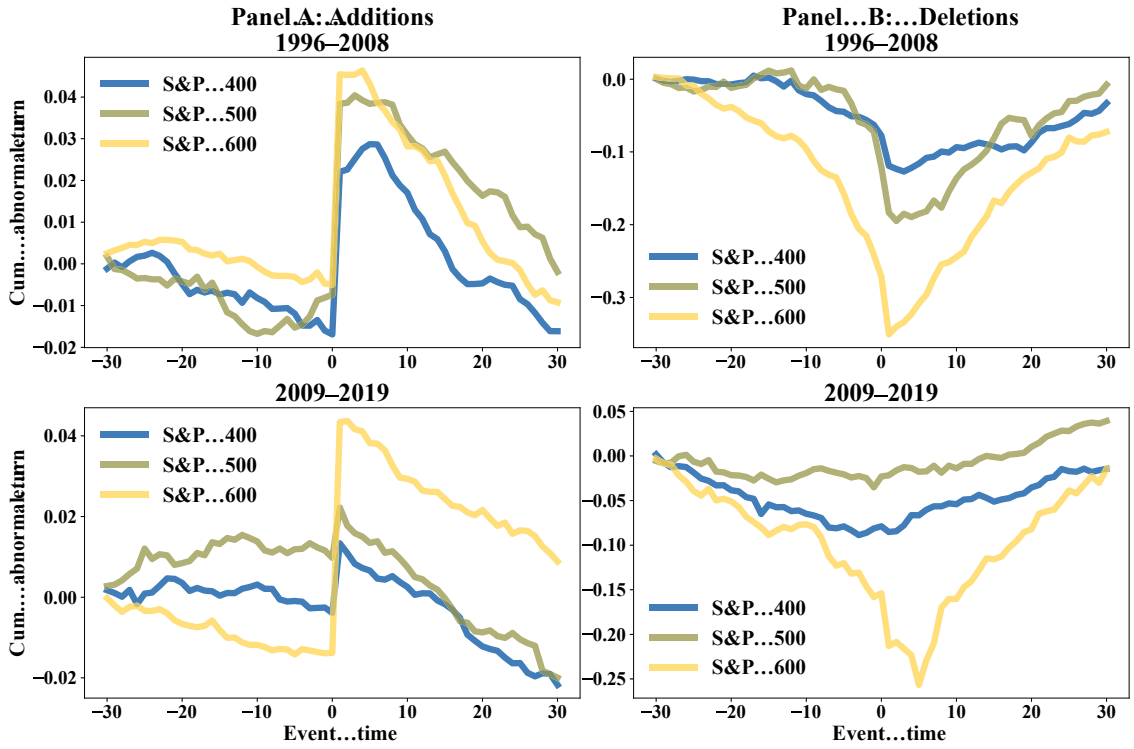

Figure 3.6: Cumulative abnormal return of S&P index additions and deletions
This figure presents the cumulative average abnormal return associated with S&P index additions (Panel A) and deletions (Panel B). The sample is split into January 1, 1996, to December 31, 2008 (upper half) and January 1, 2009, to December 31, 2019 (lower half).

The results presented in Figure 3.7 provide further insights into the behavior of the S&P 500 in response to index changes. For the S&P 500, the index addition and deletion events are separated into pure and transfer events. For additions, we find a striking difference between the pure and transfer events, particularly in the 1996–2008 period. In the 2009–2019 period, we find that index transfers have virtually no abnormal price response at $t = 0$ (AD); in fact, they trend down in the period after the announcement. Furthermore, for pure additions in the same 2009–2019 period, the abnormal price response almost entirely disappears after 30 days, which is in contrast with the 1996–2008 period, where it remains significant. Panel B shows a significant change in behavior for pure deletions. In 1996–2008, significant AD negative responses are followed by a reversion. However, in 2009–2019, the cumulative return consistently decreases in the lead-up to the announcement and revert post-announcement. This reversionary behavior could indicate that the market is better able to anticipate S&P index deletions ahead of the official announcements. The reversion could also be evidence that market-capitalization-weighted benchmarks are momentum-like in nature, with the index selling out of losing stock positions (deletions), which appear to revert once the index has sold them. Arnott, Brightman, Kalesnik and Wu (2022) argue similarly.

However, whether these stocks would have still reverted if an index change was not announced is an open question.

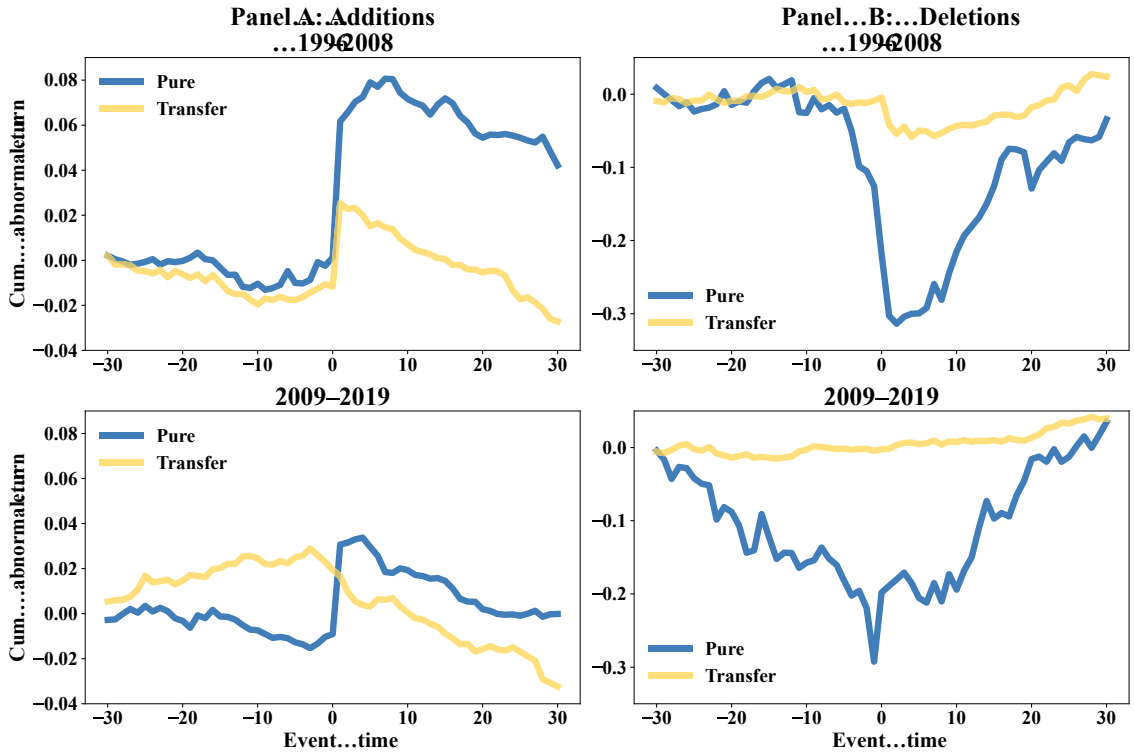

Figure 3.7: Cumulative abnormal return for S&P 500 index additions and deletions conditioned on S&P 400 transfers
This figure presents the cumulative average abnormal return associated with S&P 500 index additions (Panel A) and deletions (Panel B). The sample is split into January 1, 1996, to December 31, 2007, and January 1, 2008, to December 31, 2019. The sample is also split into stocks that are transferred between S&P indexes (transfer) and stocks that are either deleted from the S&P 1500 universe or added from outside the S&P 1500 universe (pure).

Although the results clearly show that the S&P index effect is alive for pure addition events, we cannot draw causality from this. We cannot determine if the change in the internal transfer behavior from S&P affected the average abnormal returns. It is plausible that investors have recognized this price response change and adjusted their own behavior accordingly. However, I conclude that "the S&P index effect is dead" claim is inaccurate and overstated. There are still statistically, and economically significant abnormal returns associated with pure additions from outside the S&P 1500 across all three S&P indexes. Rather, I show that "S&P index transfers are now less informationally relevant."

3.3.3 Information dynamics

S&P index changes are ultimately informational events that market participants must process. The degree to which index announcements are perceived as informationally relevant influences how market participants process these events. The informational content of S&P index events can be measured using the concept of entropy from information theory. Entropy is a measure of the average information contained in the outcomes of a random variable. To calculate entropy for index changes, AD abnormal returns are encoded in a binary fashion. Positive abnormal returns are encoded as one and negative abnormal returns as zero. The entropy formula can then be applied across different samples:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b(P(x_i)), \tag{3.6}$$

where n is the number of outcomes (index announcements), b is the base of the logarithm used, x_i are the possible outcomes (positive or negative abnormal returns), and $P(x_i)$ is the probability of each outcome. $P(x_i)$ is calculated using the chosen sample of events. As binary encoding is used, $b = 2$ is the base, and the maximum possible entropy value is $H(X) = \log_2(2) = 1$. The closer $H(X)$ is to one, the less information is contained in the sample of events. $H(X) = \log_2(2) = 1$ can be interpreted as sampling from a distribution of abnormal returns having a 50% probability of a positive value and 50% probability of a negative value. Thus, no new information is gained from each draw from this distribution.

Table 3.4 presents the entropy of the AD abnormal returns partitioned across several categories. For pure additions across all indexes, the entropy has decreased between the two sub-periods. In particular, for the S&P 600, the entropy decreases from 0.502 to 0.222, indicating that the abnormal returns of S&P 600 pure additions contain more information in the 2009–2019 period than in the 1996–2008 period. These findings are consistent with Table 3.2, which reports an increase in abnormal returns for this category between the two sub-periods. The entropy of transfers across all three indexes has increased close to one in the 2009–2019 period, suggesting that these events now contain very little information. This result aligns with the previous finding of statistically insignificant abnormal returns for transfer events.

Table 3.4: Information entropy of announcement day abnormal returns
This table presents the Shannon entropy of the AD abnormal returns of S&P index addition and deletion events. Every AD abnormal return is encoded as positive or negative. Within each sub-sample, the distribution of encoded AD abnormal returns is used to calculate the Shannon entropy, which is reported in this table. The maximum Shannon entropy value here is $\log_2(2) = 1$. The closer to zero the entropy is, the more information content is present in the sample. The Base Sample runs from January 1, 1996, through December 31, 2019. Stocks are classified as those that enter or leave the S&P 1500 universe (pure) or stocks that internally move between indexes (transfer). The sample is divided into two sub-periods: January 1, 1996, through December 31, 2008, and January 1, 2009, through December 31, 2019.

		Additions				Deletions			
	Sample	All	S&P 500	S&P 400	S&P 600	All	S&P 500	S&P 400	S&P 600
1996–2008	All	0.506	0.409	0.566	0.502	0.769	0.495	0.808	0.820
	Pure	0.441	0.227	0.334	0.502	0.771	0.503	0.737	0.820
	Transfer	0.685	0.496	0.827		0.762	0.485	0.896	
2009–2019	All	0.635	0.943	0.798	0.222	0.924	1.000	0.978	0.577
	Pure	0.241	0.341	0.260	0.222	0.667	0.881	0.877	0.577
	Transfer	0.982	0.997	0.942		0.997	1.000	0.990	

For deletions that result in the removal of stocks from the S&P 1500 universe, the entropy of S&P 400 and S&P 500 deletions has increased, indicating less information content in these events. However, the entropy of deletions of S&P 600 has increased, suggesting more information content in these events. Overall, Table 3.4 supports the finding that index transfers have become less informationally relevant in recent years. At the same time, stocks added to or deleted from the S&P 1500 universe are still informationally important events for market participants.

3.4 Horse-racing drivers of index changes

Having established that there has indeed been a mutation in the S&P index effect, a regression framework is used to explain the cross-sectional variation in the abnormal return responses. Proxies for informed trading in the days preceding the announcement of index changes are derived from options-based trading variables. A measure of relative passive ownership between large and mid-capitalization stocks is used as a proxy for the shifting relevance of the S&P 400 over the S&P 500 index.

3.4.1 Informed trading of options

The literature on options trading variables before corporate events has extensively covered the use of these variables in event studies.[7] The consensus in the literature is that informed options trading before scheduled and unscheduled corporate events

[7]See Augustin and Subrahmanyam (2020) for a detailed review.

is pervasive. This is due to the ease of obtaining leveraged positions in the options markets and lower limits to expressing short views when compared to the listed equity markets. In addition, expressing a view on outcomes over different time horizons and event probabilities is an attractive feature of options markets. It is, therefore, not surprising that informed trading in the options market occurs before corporate events that can impact a company's share price. The nature of this informed trading is of particular interest. In some cases, it is illegal, with insiders trading on private inside information and using the options markets to mask their behavior. The U.S. Securities and Exchange Commission (SEC) charged two individuals (one being an S&P employee) on September 21, 2020, with insider trading (SEC, 2020, release no. 2020-217), alleging that the named individuals "*repeatedly purchased call or put options of publicly traded companies hours before public announcements that those companies would be added to or removed from a popular stock market index.*" This case was well publicized, but this is likely not the only informed trading taking place before the announcement of index events.

Abnormal options activity before index announcements is not limited to insider trading. Although insider trading is one source for such activity, the S&P 500 index effect is a well-known phenomenon where significant trading profits can be generated by accurately predicting index changes. Moreover, while subjectivity is involved in determining the index changes, there are still predictable characteristics associated with them. For instance, candidates for index inclusion must meet specific liquidity and capitalization requirements. Therefore, it is possible to compile a list of potential index candidates that meet current index inclusion criteria but are not yet in the index and then rank them by market capitalization. Informed traders who can predict these index changes may use the options market to express their position.

3.4.2 Passive ETF ownership

The amount of assets invested in passive vehicles such as index-tracking mutual funds and ETFs has substantially grown from 1996 to 2019. Compounded capital appreciation of assets invested in the stock market and a significant reallocation from active toward passive investment strategies have driven this growth. The increased prevalence of passive investment activity has led to a new stream of literature focusing on the impact of passive investing on stock markets. Although this is a relatively new area of research, there is mixed evidence on the implications of the growth in passive assets on financial markets. Additionally, the operation of passive investment strategies has received less attention in academic literature.

A typical passive investment strategy involves a pool of assets attached to a benchmark index provided by a separate vendor such as S&P, FTSE Russell, or MSCI. The passive investment vehicle pays a licensing fee to the vendor to market a passive strategy benchmarked to their index. Then, the manager of the passive strategy seeks to minimize the fund's tracking error to the official benchmark maintained by the vendor. As the start-of-day weights associated with the benchmark are disseminated daily, only a limited number of activities can cause the passive vehicle's performance to deviate from the benchmark's performance. One is associated with the provider's mechanism for adding and removing companies from their benchmark. Each provider typically has their own method for implementing index changes, but a common theme is that the changes are announced (the AD) ahead of the date the change will be made (the ED).

The periods around index changes are of heightened importance due to the required increase in trading for passive vehicles around these periods. These vehicles are required to mechanically rebalance to maintain a low tracking error to their benchmark, which often occurs in a very limited period. Historically, this type of behavior canonically drives part of the index effect. If there is a known date in the future when uninformed passive assets will be required to purchase a stock, arbitrage traders could buy the stock ahead of this date and then demand a higher price for providing the stock on the given future date.

Although the absolute level of passive assets has increased, what is more important to the index effect is the relative level of passive ownership across different market capitalizations of stocks. Figure 3.8 presents the normalized passive ETF ownership for stock rank market capitalizations, plotting each year on a separate line. Over time, a clear trend is observed. Before 2005, large- and mega-capitalization stocks had relatively higher levels of passive ETF ownership than smaller capitalization stocks. After 2005, this trend inverted, with stocks in the 900–1500 market capitalization rank having relatively higher levels of passive ETF ownership. I construct an annual measure of the relative change in passive ETF ownership levels between the S&P 500 and S&P 400 by considering the difference in the average passive ETF ownership by stock ranks 1–500 and stock ranks 501–900.

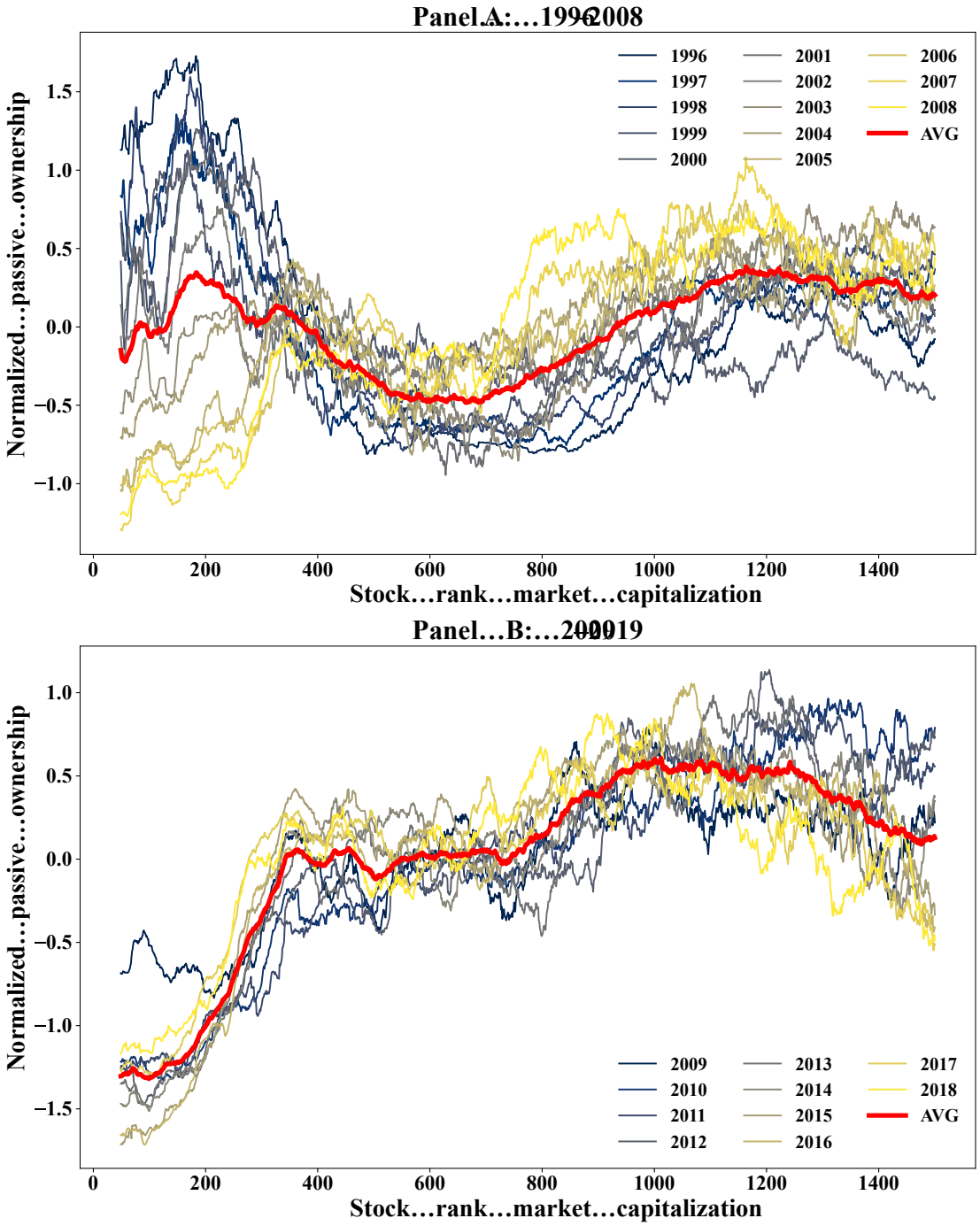

Figure 3.8: Annual trends in the passive ETF ownership of U.S. stocks sorted by market capitalization

This figure presents the annual normalized passive ETF ownership for the top 1500 stocks in the CRSP universe sorted by market capitalization. Each month, in each year, all stocks in the CRSP universe are ranked by market capitalization (one corresponding to the largest market capitalization). For each year and each stock, I take the average passive ETF ownership across each month. Each year, the passive ETF ownership is normalized by the annual mean and standard deviation of passive ETF ownership. The red line represents the cross-sectional average for each stock rank market capitalization across the years represented in each panel. A rolling 50 stock rank average in each year is reported for smoothing purposes.

3.4.3 Regressions

Cross-sectional variation in AD abnormal returns is explained using a general panel regression model:

$$AR_{i,0} = \alpha + \sum_{k=1} \beta_k AV_{i,t[-20,-1]} + \beta_p PASSOWN_{t-1} + \beta_A \text{PURE ADD}_{i,t} \qquad (3.7)$$
$$+\beta_D \text{PURE DELETE}_{i,t} + \sum_{j=1} \beta_j CONTROL_{i,t} + \epsilon_{i,t},$$

where $AR_{i,0}$ is the abnormal return for stock i on the event day ($t = 0$), $AV_{i,t[-20,-1]}$ is the average of the abnormal value of the variables of interest in the pre-event period ($t = -20$ to $t = -1$), $PASSOWN_{t-1}$ is the difference in the average passive ownership between U.S. stocks sorted by market capitalization in the 1–500 and 501–900 groups as at the end of December of the year before the event, PURE $\text{ADD}_{i,t}$ is a dummy variable equal to one if the stock was added from outside the S&P 1500 universe and zero if the stock was transferred upward from one of the other S&P indexes, PURE $\text{DELETE}_{i,t}$ is a dummy variable equal to one if the stock was deleted from the S&P 1500 universe or zero if the stock was transferred downward to one of the other S&P indexes, and $CONTROL_{i,t}$ is a set of control variables.

The independent variables of interest (denoted by $AV_{i,t}$ in Eq. (3.7)) are the abnormal CPIV spread, abnormal call options volume share (CVOLSHR), abnormal stock trading volume share (ABSVOLSHARE), and change in passive ownership trends ($PASSOWN_t$). The control variables used are book-to-market (BtM), idiosyncratic stock volatility (STD), log market capitalization (SIZE), stock price 20 days prior to the event (PRC), stock volume share (VOLSHARE), the number of analysts covering the stock (NUMANAL), and short interest ratio (SHORTINT). The full variable definitions are presented in Appendix 3.5.

Table 3.5 presents the summary statistics for the independent variables used in the Regression Sample of additions and deletions for the three indexes. For each index, I present results for the full sample and the pure/transfer sub-samples. As I have established, an aggregation using the full sample misses an important distinction between internal index changes (index promotions and demotions) and external additions and deletions (pure). By splitting the full sample into the two sub-samples, we find consistently different sample statistics across all three indexes. For example, in Panel B for the S&P 400, the average abnormal return on the AD is 3.08%. However, for the pure sub-sample, the average abnormal return is 5.09%; for the transfer sub-sample, it is 0.94%. This is a stark difference and can lead to very

different conclusions around the economic significance of the AD abnormal returns. Similarly, in Panel A for the S&P 500, the call option volume share for the deletion samples varies considerably.

Table 3.5: Summary statistics of the regression sample
This table presents the summary statistics for the S&P index additions and deletions Regression Sample. The Regression Sample runs from January 1, 1996, to December 31, 2019. μ is the sample mean, and σ is the sample standard deviation. Panel A presents results for the S&P 500. Panel B presents results for the S&P 400. Panel C presents results for the S&P 600. For the S&P 600, statistics are reported only for the Pure Sample, as transfer events are not relevant to the S&P 600 in this study. Full variable definitions are presented in Appendix 3.5.

Panel A: S&P 500												
	Additions						Deletions					
	Full Sample		Pure		Transfer		Full Sample		Pure		Transfer	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
AR_0 (%)	3.23	4.30	5.24	4.31	2.02	3.82	-0.90	10.52	-3.34	20.84	-0.12	3.10
$AR_{[0,10]}$ (%)	2.09	9.49	6.02	9.45	-0.28	8.71	2.73	23.81	7.52	46.49	1.21	8.49
ABVOLSHARE (%)	0.01	0.41	-0.03	0.43	0.03	0.39	0.38	0.96	0.95	1.71	0.20	0.41
CPIV	-0.08	4.09	0.16	4.53	-0.22	3.81	0.24	11.34	0.24	23.08	0.24	2.35
CVOLSHR (%)	0.21	1.50	0.22	2.25	0.20	0.75	0.26	1.05	0.67	1.50	0.14	0.83
PASSOWN (%)	-0.36	0.68	-0.45	0.70	-0.31	0.66	-0.96	0.62	-0.35	0.60	-1.15	0.49
BtM	0.34	0.31	0.42	0.38	0.29	0.26	0.91	0.67	1.19	0.73	0.82	0.63
NUMANAL	11.06	9.46	13.06	9.71	9.86	9.12	9.63	8.72	6.77	7.12	10.54	9.02
PRC	70.85	57.87	60.41	55.82	77.14	58.31	19.94	17.56	10.31	12.67	22.99	17.84
SI (%)	2.91	4.54	1.94	3.03	3.50	5.17	9.40	9.09	5.94	7.01	10.50	9.43
SIZE	8.95	0.63	9.22	0.79	8.78	0.44	8.03	0.82	7.89	1.44	8.07	0.50
STD (%)	2.58	1.77	2.61	2.08	2.57	1.56	4.39	5.69	9.68	9.42	2.72	1.87
VOLSHARE (%)	1.42	1.24	1.54	1.39	1.35	1.14	2.38	1.74	2.34	2.21	2.39	1.58
Panel B: S&P 400												
	Additions						Deletions					
	Full Sample		Pure		Transfer		Full Sample		Pure		Transfer	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
AR_0 (%)	3.08	4.62	5.09	4.01	0.94	4.26	-0.68	6.81	-5.47	9.25	1.02	4.73
$AR_{[0,10]}$ (%)	1.52	10.71	4.59	10.52	-1.76	9.94	3.45	15.39	4.96	23.07	2.92	11.69
ABVOLSHARE (%)	0.00	0.40	0.00	0.43	0.01	0.38	0.22	0.54	0.32	0.73	0.18	0.45
CPIV	-0.27	3.56	-0.34	4.08	-0.19	2.90	1.21	7.34	0.96	10.65	1.29	5.83
CVOLSHR (%)	0.16	1.04	0.14	1.01	0.18	1.06	0.09	0.91	-0.09	0.97	0.15	0.89
PASSOWN (%)	-0.38	0.72	-0.27	0.69	-0.49	0.73	-0.91	0.67	-0.36	0.60	-1.11	0.59
BtM	0.33	0.25	0.32	0.26	0.34	0.23	0.95	0.81	1.02	0.89	0.93	0.79
NUMANAL	6.70	6.51	8.04	6.92	5.28	5.72	8.20	7.83	7.52	7.36	8.44	8.02
PRC	52.32	38.71	43.34	24.89	61.89	47.60	15.11	14.59	12.40	22.52	16.07	10.50
SI (%)	11.98	18.10	11.16	16.42	12.86	19.74	27.63	37.37	25.07	37.50	28.54	37.52
SIZE	7.69	0.55	7.65	0.67	7.72	0.39	7.09	0.63	7.06	0.92	7.11	0.50
STD (%)	2.71	1.65	2.97	1.73	2.43	1.51	4.11	3.29	6.83	4.62	3.15	1.93
VOLSHARE (%)	1.56	0.95	1.63	1.02	1.48	0.87	2.32	1.40	2.55	1.81	2.23	1.22

Panel C: S&P 600				
	Additions		Deletions	
	Pure		Pure	
	μ	σ	μ	σ
AR_0 (%)	5.46	4.68	-2.19	8.49
$AR_{[0,10]}$ (%)	3.44	11.65	7.43	27.93
ABVOLSHARE (%)	-0.04	0.49	0.31	0.97
CPIV	0.02	5.30	4.75	13.18
CVOLSHR (%)	-0.01	0.94	-0.15	0.49
PASSOWN (%)	-0.32	0.70	-0.49	0.69
BtM	0.35	0.30	1.20	1.69
NUMANAL	4.86	4.82	5.72	6.47
PRC	26.70	12.30	21.63	36.02
SI (%)	30.86	45.74	58.73	73.03
SIZE	6.47	0.58	6.27	1.03
STD (%)	3.16	1.68	7.82	5.77
VOLSHARE (%)	1.84	1.16	2.12	1.46

Table 3.6 shows the Pearson correlation coefficients between the regression variables and the AD abnormal returns within the additions (upper right) and deletions (lower left). The largest correlated variables for the additions are the price 20 days prior to the index announcement (PRC) and the idiosyncratic volatility (STD). The largest correlated variables for the deletions are the call option volume share (CVOLSHR) and PRC.

Table 3.6: Correlation in the regression sample
This table presents the Pearson correlation coefficients between variables in the regression variable set. The Regression Sample used runs from January 1, 1996, to December 31, 2019. The upper right triangular values correspond to the correlations for additions. The lower left triangular values correspond to the correlations for deletions. Full variable definitions are presented in Appendix 3.5.

		Additions										
		AR_0 (%)	SIZE	BtM	PRC	NUMANAL	VOLSHARE	STD	SHORTINT	ABSVOLSHARE	CPIV	CVOLSHR
Deletions	AR_0 (%)		-19.4	4.5	-23.8	-4.7	4.5	23	2.0	-6.1	0.4	-6.4
	SIZE	7.5		-8.0	46.2	38.1	-10.2	-19.3	-32.2	0.1	4.7	0.8
	BtM	8.8	-17.9		-15.6	4.5	-14.3	-16.5	-0.7	0.0	-3.0	4.0
	PRC	15.1	32.1	-25		16.2	4.6	-10.2	-0.9	2.7	1.6	3.7
	NUMANAL	5.2	28.6	12.8	-9.4		2.4	-8.5	-7.9	-2.2	2.8	0.0
	VOLSHARE	13.8	1.3	15.5	-19	14.6		32.3	28.7	-7.3	5.8	-5.9
	STD	5.9	-10.1	13.7	-25.3	0.7	15.3		-1.8	23.6	-3.8	4.9
	SHORTINT	7.0	-40.3	10.3	-5.7	-14.7	22.6	16.9		5.0	-0.9	7.6
	ABSVOLSHARE	5.0	13.5	1.9	-12.4	3.6	-10.6	44.0	6.1		-4.7	50.4
	CPIV	5.3	6.2	-8.4	0.8	-1.1	8.2	16.9	-5.5	-14.1		-1.8
	CVOLSHR	15.5	0.1	0.2	-3.1	0.8	-10.4	16.7	19.5	64.8	-9.1	

Table 3.7 presents the results of applying Eq. (3.7) across a series of different index change panel datasets. The Full Sample regressions use a panel of addition and deletion events, with type fixed effects on the addition and deletion categories.

The Full Sample regression is run on the three independent index panels, and an additional regression (denoted as All) using events from all three indexes is included. A statistically significant loading on PURE ADD for S&P 500 and S&P 400 additions establishes the relation between the index effect and membership in a lower index. Stocks that are added to the S&P 500 or S&P 400 from outside the S&P 1500 universe experience on average higher AD abnormal returns of 3.55% and 3.38%, respectively, when compared with stocks that are promoted from a lower capitalization index. A similar pattern occurs for PURE DELETE events. Stocks completely deleted from the S&P 1500 experience significantly more negative AD abnormal returns than stocks demoted to lower capitalization indexes. One explanation for this result is investor attention. Historically, the S&P 500 received the most attention in mainstream media, but more recently, the S&P 400 and S&P 600 have gained more attention. Now, an increase in investor attention toward a stock drives an abnormal price response when the stock is first added to the broader S&P 1500 universe.

Table 3.7: Options implied volatility measures, passive ownership, and S&P 1500 index additions and deletions announcement day abnormal returns
This table presents the coefficient estimates from cross-sectional regressions of the S&P 500, S&P 400, and S&P 600 index additions and deletions AD abnormal stock returns on a set of IV, passive ownership, and risk-based independent variables. Regressions follow Eq. (3.7). Abnormal value market models are calibrated in the window between $t = -150$ and $t = -31$. Index additions/deletions type fixed effects are used, and standard errors are clustered by firm. The sample runs from January 1, 1996, to December 31, 2019. t-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Full Sample				Additions				Deletions		
	S&P 500	S&P 400	S&P 600	All	S&P 500	S&P 400	S&P 600	All	S&P 500	S&P 400	All
CPIV	-0.021	0.001	-0.087*	-0.048	-0.108**	0.116	-0.032	-0.020	0.021	0.057	-0.078
	(-0.16)	(0.01)	(-1.69)	(-0.85)	(-2.01)	(1.61)	(-0.64)	(-0.56)	(0.10)	(0.54)	(-0.71)
CVOLSHR	-0.547*	-0.077	0.017	-0.138	-0.390***	-0.554***	-0.247	-0.403***	-0.618	1.575*	1.703
	(-1.88)	(-0.27)	(0.07)	(-0.89)	(-3.02)	(-2.61)	(-0.93)	(-3.86)	(-0.41)	(1.74)	(1.49)
ABSVOLSHARE	0.575	-0.302	-1.955***	-0.900	-0.721	-0.451	-1.053**	-0.671*	2.102	1.252	-1.426
	(0.23)	(-0.50)	(-3.56)	(-0.86)	(-0.94)	(-0.78)	(-2.00)	(-1.95)	(0.56)	(0.93)	(-0.67)
PASSOWN	0.340	1.260***	-0.746**	0.261	1.832***	0.933***	-0.640*	0.628***	-4.184*	1.177	-2.645**
	(0.50)	(3.51)	(-2.00)	(0.95)	(4.63)	(3.22)	(-1.78)	(3.28)	(-1.77)	(0.94)	(-2.48)
PURE DELETE	-11.022***	-7.936***		-7.405***					-8.225**	-4.959***	-5.635***
	(-4.48)	(-4.30)		(-6.00)					(-2.45)	(-2.81)	(-3.69)
PURE ADD	3.130***	3.801***		3.479***	3.548***	3.376***		3.436***			
	(5.30)	(8.21)		(11.42)	(6.82)	(7.41)		(11.74)			
SHORTINT	-0.042	-0.007	-0.003	-0.002	-0.043	0.009	0.001	-0.000	-0.113	-0.003	-0.015
	(-0.78)	(-0.70)	(-0.59)	(-0.41)	(-1.02)	(0.73)	(0.23)	(-0.06)	(-0.89)	(-0.18)	(-0.94)
STD	0.976***	0.048	0.417**	0.597**	0.464**	0.666***	0.597***	0.582***	1.053***	-0.672**	0.721*
	(3.10)	(0.28)	(2.42)	(2.56)	(2.56)	(3.52)	(3.06)	(5.29)	(2.90)	(-2.29)	(1.84)
BtM	1.138	1.197**	1.275***	1.280**	-0.050	0.040	1.833*	0.872*	0.852	2.043**	1.122*
	(0.88)	(2.40)	(2.59)	(2.48)	(-0.08)	(0.06)	(1.93)	(1.87)	(0.59)	(2.50)	(1.79)
SIZE	-0.335	1.126**	-0.292	0.176	-0.320	0.694	-0.135	0.232	-2.281	2.270*	-1.088
	(-0.36)	(2.32)	(-0.53)	(1.03)	(-0.71)	(1.46)	(-0.24)	(1.45)	(-0.76)	(1.79)	(-1.15)
PRC	-0.001	-0.006	0.028	-0.004	0.002	-0.014**	-0.022	-0.007**	0.103	0.010	0.109***
	(-0.09)	(-0.93)	(1.10)	(-1.00)	(0.49)	(-2.01)	(-1.16)	(-2.31)	(1.36)	(0.24)	(3.57)
VOLSHARE	0.141	0.657**	-0.577**	-0.001	-0.002	0.238	-0.702***	-0.176	0.655	0.613	0.480
	(0.54)	(2.33)	(-2.19)	(-0.00)	(-0.01)	(0.82)	(-2.66)	(-1.27)	(1.25)	(1.34)	(1.23)
NUMANAL	0.018	-0.031	-0.111**	-0.021	0.016	-0.011	-0.094**	-0.010	0.026	-0.036	-0.027
	(0.73)	(-1.01)	(-2.13)	(-1.06)	(0.68)	(-0.37)	(-2.02)	(-0.58)	(0.34)	(-0.44)	(-0.45)
Intercept	1.401	-7.936**	5.658	-1.609	4.132	-5.035	5.852*	-1.016	6.566	-15.274*	0.033
	(0.20)	(-2.18)	(1.64)	(-1.02)	(1.08)	(-1.35)	(1.66)	(-0.79)	(0.33)	(-1.93)	(0.01)
N	422	514	419	1355	314	403	394	1111	108	111	244
Adj. R^2	0.270	0.252	0.090	0.178	0.331	0.329	0.082	0.230	0.366	0.360	0.235
Type FE	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No

Figure 3.8 shows how in recent years, the relative passive ownership of mid- and small-capitalization companies is higher than that of large- and mega-capitalization companies, despite the S&P 500 having substantially more passive assets tracking it than the S&P 400 or S&P 600. This shift may be attributable to the larger number of large-capitalization active mutual funds compared with mid- and small-capitalization managers, which comprise a relatively larger proportion of the large- and mega-capitalization share registries. A statistically significant positive loading on the PASSOWN variable is found for S&P 500 and S&P 400 additions. This variable is measured as a time-series variable from the previous year for each index event and has a constant value for events occurring in the same year. It measures the relative importance of the trend of PASSOWN on abnormal returns. For increases in PASSOWN, S&P 500 and S&P 400 additions experience larger AD abnormal returns. This relation captures the fact that abnormal returns, in general, were higher in the earlier periods of the sample, which also correspond to higher values of PASSOWN. As PASSOWN has become more negative (i.e., the relative passive ownership of S&P 400 to S&P 500 companies has increased), the AD abnormal returns have also decreased. The S&P 400 effect is roughly half the effect for the S&P 500, reflecting the relative imbalance between pure additions and transfers for the S&P 400 compared with the S&P 500.

The regression results presented in Table 3.7 show several statistically significant loadings on CPIV and CVOLSHR. Significant positive loadings on CPIV for additions to the S&P 500 indicate how changes in the underlying IV surfaces influence AD abnormal returns. Abnormal values of CPIV suggest that investors expect an increase in the stock price and demand more ATM call options, which drives an increase in the price of ATM call options. For additions, I find statistically significant negative loadings on abnormal CVOLSHR. There is generally a smaller AD abnormal return for higher levels of abnormal CVOLSHR. Given the increased abnormal CVOLSHR ahead of the event, this result suggests that abnormal options trading leading up to the index announcement event results in weaker observed index effects.

Table 3.7 also presents the results for index deletions. In the All Sample, I include S&P 600 deletions, noting that after accounting for available data, the sample for S&P 600 deletions only consists of 18 events. The All Sample has a statistically significant negative loading on PASSOWN. The sign of this loading is consistent with the result for the index additions, suggesting that the changing structure of passive ownership has also reduced the negative abnormal return responses when stocks are deleted from the S&P indexes.

Table 3.8: Options implied volatility measures, passive ownership, and S&P 1500 index additions announcement day abnormal returns
This table presents the coefficient estimates from cross-sectional regressions of the S&P 500, S&P 400, and S&P 600 index additions and deletions AD abnormal stock returns on a set of IV, passive ownership, and risk-based independent variables. Regressions follow Eq. (3.8). Abnormal value market models are calibrated using the window between $t = -150$ and $t = -31$. Standard errors are clustered by firm. The sample runs from January 1, 1996, to December 31, 2019. t-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	S&P 500		S&P 400	
	Pure	Transfer	Pure	Transfer
CPIV	-0.040	-0.133**	0.146**	0.048
	(-0.39)	(-2.19)	(2.09)	(0.32)
CVOLSHR	-0.354	-0.679*	-0.517**	-0.467
	(-1.58)	(-1.67)	(-2.52)	(-1.27)
ABSVOLSHARE	-0.442	-0.761	-0.233	-0.501
	(-0.28)	(-0.90)	(-0.39)	(-0.40)
PASSOWN	1.196**	2.319***	-0.029	1.591***
	(2.08)	(4.01)	(-0.07)	(4.12)
SHORTINT	-0.225**	0.010	-0.033**	0.047***
	(-2.11)	(0.19)	(-2.20)	(2.89)
STD	0.932***	0.126	0.702***	0.590
	(3.66)	(0.50)	(3.58)	(1.54)
BtM	-0.075	0.165	-0.400	0.187
	(-0.07)	(0.19)	(-0.43)	(0.20)
SIZE	-0.453	-0.762	-0.215	2.356*
	(-0.80)	(-1.04)	(-0.57)	(1.97)
PRC	0.013*	-0.001	0.013	-0.027***
	(1.83)	(-0.14)	(1.25)	(-3.68)
VOLSHARE	-0.266	-0.037	0.068	0.562
	(-0.71)	(-0.17)	(0.19)	(1.33)
NUMANAL	0.064	-0.015	-0.036	-0.008
	(1.53)	(-0.49)	(-1.00)	(-0.17)
Intercept	6.875	9.378	4.860	-17.574*
	(1.36)	(1.53)	(1.58)	(-1.81)
N	118	196	208	195
Adj. R^2	0.302	0.270	0.164	0.250

Table 3.8 splits the addition sample into pure additions and transfer additions and runs the following regression on these sub-samples:

$$AR_{i,0} = \alpha + \sum_{k=1} \beta_k AV_{i,t[-20,-1]} + \beta_p PASSOWN_{t-1} + \sum_{j=1} \beta_j CONTROL_{i,t} + \epsilon_{i,t}. \quad (3.8)$$

The most striking result is the difference in loadings between the pure and transfer samples. In particular, the PASSOWN variable is significantly stronger for the transfer samples of the S&P 500 and S&P 400. This result can be interpreted as follows: the relative effect of the change in passive ownership between mid- and large-capitalization stocks results in lower abnormal returns for index transfers than for pure additions. This result is supportive of the view that the increasing importance and attractiveness of the S&P 400 and S&P 600 indexes have meant that market participants are agnostic to which S&P index a stock is added into, as long it is added into the broader S&P 1500 index universe and passive investors are required to purchase the stock.

3.4.4 Economic explanations

The previous results establish that index transfer events are no longer informative events that market participants respond to. In this section, I present several potential economic mechanisms for the declining informativeness of index transfer events.

When a stock is promoted from the S&P 400 to the S&P 500, passive asset managers must either buy or sell it, depending on which index they track. The total value of the assets that passively track each index determines the required net buying and selling. For instance, suppose stock A has a weight of 10% in the S&P 400 and will be promoted to the S&P 500, where its new weight will be 1%. If the S&P 500 has $1000 of assets tracking it and the S&P 400 has $100 of assets tracking it, then the net buying and selling required is zero. Typically, the weight of the stock in the index becomes smaller when a stock is promoted from a lower capitalization index to a higher capitalization index. In the 2000s and earlier, the S&P 400 and S&P 600 had significantly lower levels of assets passively following them. Thus, being promoted to the S&P 500 resulted in significant net buying demand, which increased the share price. However, as the amount of passive assets tracking the S&P 400 and S&P 600 increased, the net buying demand resulting from an upward transfer (from the S&P 400 to the S&P 500) has become smaller. This supply–demand mechanism can reduce the buying–selling effect required around index rebalances across market capitalizations.

Another important consideration is the management of portfolios by large passive investment managers such as State Street, Vanguard, and BlackRock, also colloquially known as the "Big Three." These managers offer managed funds and passive ETF products that track all three S&P indexes. Given their large scale of management, it is likely that passive asset managers can internally cross significant

amounts of stock between different funds when rebalancing their passive mandates. As a result, these funds do not have to trade shares on the open market but can trade with other internal funds, paying no brokerage fees to trade the required shares. This crossing ability significantly reduces the on-market buying and selling activity in index transfer events, which can result in smaller price responses.

It is also possible that index transfers have become less informative events for companies because they are easier to predict than pure additions. Suppose a company has already met certain internal (non-disclosed) S&P metrics required for inclusion in the broader S&P 1500 universe. In that case, a stock's market capitalization and operating industry may become the primary determinants of whether the stock will be promoted from a lower index to a higher one. This greater predictability can make it easier for index arbitrageurs and passive index fund managers to anticipate these index transfers more accurately prior to their announcement. Thus, the index effect may have declined for transfer events, as potential effects have already been priced-in ahead of the transfer announcement.

3.5 Conclusion

Including a stock in the S&P 500 is broadly accepted as a positive event for the company. However, recent evidence suggests that inclusion into the S&P 500 no longer positively affects a company's stock price, leading to claims of the death of the S&P index effect. This supposed death of the index effect has occurred while the assets allocated to passive index tracking strategies reached an all-time high. I investigate the S&P index effect across the different U.S. index market capitalization classes by examining the origin and destination of index additions and deletions. I conclude that the statement "the S&P index effect is dead" is too strong. Instead, the S&P index effect has reduced for companies moving from a lower capitalization index into a higher capitalization index and remains persistent but statistically significant for additions from outside the current S&P 1500 index universe.

To address the conflict of an increase in passively managed assets alongside a supposed decline in the index effect, I use a regression framework to demonstrate that the relative increase in the passive ownership of mid-capitalization companies to large-capitalization companies helps explain why index transfers are now less informative and experience no abnormal price response on the AD of index changes. My results suggest that even with the growth in passive assets, the market is effectively digesting index changes without significant distortions to the underlying companies. The market efficiently distinguishes between index announcements that

contain company-relevant information (such as an addition from outside the S&P 1500) and announcements that should not substantially impact the share price (such as movement between S&P indexes). For passive asset managers, particularly of S&P 400 and S&P 600 strategies, the results of this chapter show that there is still a benefit to allocating tracking error budget to hold potential index inclusion stocks that are not already in another S&P index.

Appendix 3.1. Variable definitions

Options measures

- IV (implied volatility): I measure IV as the EW IV of all valid options for a given call-put moneyness group. I take all options with best bid greater than \$0, mid-price greater than \$0.25, open interest greater than zero, and a bid-ask spread ratio less than 50%. I define option moneyness for calls and puts as:

$$Calls = IV_C = \begin{cases} IV_{C,ATM} = & \left|\log\left(\frac{P_t}{K_t}\right)\right| < 0.1, \\ IV_{C,ITM} = & \log\left(\frac{P_t}{K_t}\right) > 0.1, \\ IV_{C,OTM} = & \log\left(\frac{P_t}{K_t}\right) < -0.1. \end{cases} \tag{3.9}$$

$$Puts = IV_P = \begin{cases} IV_{P,ATM} = & \left|\log\left(\frac{P_t}{K_t}\right)\right| < 0.1, \\ IV_{P,ITM} = & \log\left(\frac{P_t}{K_t}\right) < -0.1, \\ IV_{P,OTM} = & \log\left(\frac{P_t}{K_t}\right) > 0.1. \end{cases} \tag{3.10}$$

  ATM denotes at-the-money, ITM denotes in-the-money, OTM denotes out-of-the-money, P_t is the closing price at time t, and K_t is the option strike priced divided by 1,000 at time t.

- CPIV (call-put IV spread):

$$CPIV = IV_{C,ATM} - IV_{P,ATM}. \tag{3.11}$$

- CVOLSHR (call volume share):

$$CVOLSHR = \frac{CVOL_{i,t} \times 100}{SHROUT_{i,t}}, \tag{3.12}$$

  where $CVOL_{i,t}$ is the total valid traded call option volume for stock i at time t and $SHROUT_{i,t}$ is the common equity shares outstanding for stock i at time t. I multiply $CVOL_{i,t}$ by 100, as each option contract is for 100 shares in the underlying stock.

Control measures

- BtM (book-to-market): I follow Fama and French (1992) and compute a firm's book-to-market ratio in month t using the market value of its equity at the end of December in the previous year and the book value of common equity

plus balance-sheet deferred taxes minus preferred stock for the firm's latest fiscal year ending in the prior calendar year.

- NUMANAL (number of analysts): I measure the number of analysts using the Institutional Brokers' Estimate System (I/B/E/S) dataset and take the number of analysts having reported one-year-forward earnings per share (EPS) forecasts for a stock at the end of December in the previous year.

- PASSOWN (relative passive ownership): This is the difference in percentage passive ownership between CRSP stocks ranked on market capitalization in the 1–500 group and the 501–900 group.

$$PASSOWN_t = OWNR_{1-500,t} - OWNR_{501-900,t}. \tag{3.13}$$

- PRC (price): This is the stock price taken from the CRSP dataset, measured at $t = -20$ days prior to the event.

- SI (short interest): This is the ratio of short interest to shares outstanding from Compustat. I take the prevailing ratio at the end of the previous month before the event.

- SIZE (log market capitalization): I follow existing literature and measure firm size as the natural logarithm of the market value of equity (price times shares outstanding in millions of dollars) at the end of December in the previous year.

- STD (idiosyncratic volatility): I measure this as the standard deviation of the residuals from Eq. (3.3) during the event window.

- VOLSHARE (average stock volume share):

$$VOL = \frac{StockVolume_{i,t} \times 100}{SHROUT_{i,t}}. \tag{3.14}$$

- ABSVOLSHARE (abnormal stock volume share): This is the abnormal value of VOLSHARE, as measured using Eq. (3.5).

Appendix 3.2. Incremental turnover and algorithmic trading

Harris and Gurel (1986) proposed the price pressure hypothesis, suggesting that large traders all trading in the same direction induce a temporary price effect. This hypothesis has been proposed as one of the explanations for the S&P 500 index effect (Chen et al., 2004). In this study, I analyze the average incremental turnover for stock volume, call option volume, and put option volume for the samples of additions and deletions on the first trading day after the AD. The variable is defined in Section 3.2.5, following Vijh and Wang's (2022) approach.

Table 3.9 shows that in general, the incremental turnover is either not statistically different or statistically lower than the previous sub-period. In the case of additions to the S&P 600 index, there is one exception, where there is a statistically significant 73% increase in the incremental stock turnover. This finding provides some evidence that the sub-period increase in abnormal returns for additions to the S&P 600 index could be explained by the price pressure hypothesis. However, overall, we do not find a consistent change in the incremental stock turnover levels to conclude that the price pressure hypothesis explains the results.

When examining the incremental turnover of options activity, a general trend of decreasing options turnover on the AD of both additions and deletions, particularly for the S&P 500, is observed. Several reasons could explain the elevated options activity on the day of these events. The stock and options markets are inextricably linked, so elevated activity in one market will have a follow-on effect in the other. On the AD, as the market responds and the stock goes either up or down and given that we have observed typically elevated abnormal returns, it is feasible that a larger proportion of options are traded because options holders seek to crystallize profits from the abnormal move in price and de-risk their holdings.

Given the increase of algorithmic trading in recent years, and the continued move to fully electronic markets, one may have expected these index events to have become more volatile and unstable. Similarly, one potential avenue for exploiting these index events is algorithmic traders participating in the increased trading activity that is occurring.

Using the SEC MIDAS dataset, which consists of stock-day measures of various algorithmic trading variables, I measure the incremental change in these algorithmic trading variables across a period around the AD. Table 3.10 presents the average incremental change in trade-to-order and cancels-to-trade ratios across various sub-periods around the AD. I use a similar method as in Table 3.9, where I measure

the value of the trade-to-order and cancels-to-trade ratios relative to the market during the calibration window. Lower values of trade-to-order ratios are indicative of higher algorithmic trading activity, whereas higher values of cancels-to-trade ratios are indicative of higher algorithmic trading activity. The results almost unanimously show that across every index and categorization, on the AD (τ_4) and in the period between the AD and ED (τ_5), there is a substantial decrease in the algorithmic trading proxies. This suggests that on these trading days, algorithmic traders have either reduced participation in the market in these stocks relative to their normal participation or they have not increased their participation alongside the increased stock turnover that we observed in Table 3.9. The implication of this is that algorithmic trading is likely not having a significant influence on the abnormal returns which we have observed in the 2012–2019 events from the sample and that the observations are more likely driven by the mix of active investors, passive investors, and arbitrageurs who are trading in the physical stock on and around the AD (Weller, 2018).

One interesting observation from Table 3.10 is that the algorithmic trading proxies show reduced algorithmic trading behavior for index promotions and demotions as well. Although this reduction is generally not as extreme as the reduction observed for new additions and full deletions, it does suggest that a more likely explanation for the results is that algorithmic traders do not increase their activity on these high stock volume days. This result has an important implication: algorithmic traders likely are not having a large influence on these highly informative days for index events, and this is potentially a reason why these events have remained reasonably orderly despite the increased electronic execution in markets.

Table 3.9: Announcement day incremental turnover
This table presents the average incremental turnover on the AD of S&P index addition and deletions. The Base Sample used runs from January 1, 1996, through December 31, 2019. Stocks are classified as those that enter or leave the S&P 1500 universe (pure) or stocks that internally move between indexes (transfer). The sample is split into two sub-periods: January 1, 1996, through December 31, 2008, and January 1, 2009, through December 31, 2019. Abnormal returns are reported in percentages. t-statistics are reported in parentheses. t-statistics reported for difference in average abnormal returns assume unequal sample variance. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively, only for differences in values.

Panel A: Additions							
		Call Options		Put Options		Stock	
		Transfer	Pure	Transfer	Pure	Transfer	Pure
S&P 400	1996–2008 (1)	1.96	4.92	1.66	11.01	1.55	3.58
		(2.19)	(7.96)	(2.07)	(1.28)	(6.29)	(9.87)
	2009–2019 (2)	1.65	3.59	6.75	3.32	1.35	3.61
		(4.64)	(2.57)	(1.83)	(2.18)	(6.16)	(11.35)
	(2) - (1)	-0.31	-1.33	5.10	-7.69	-0.20	0.03
		(-0.34)	(-0.90)	(1.36)	(-0.95)	(-0.60)	(0.07)
S&P 500	1996–2008 (1)	6.70	9.37	4.27	9.25	5.20	6.81
		(3.78)	(4.68)	(3.54)	(3.14)	(7.79)	(5.00)
	2009–2019 (2)	1.85	3.76	1.55	2.86	2.32	4.15
		(3.94)	(6.57)	(2.65)	(4.70)	(4.36)	(8.71)
	(2) - (1)	-4.85***	-5.61***	-2.72**	-6.39**	-2.88***	-2.66*
		(-2.67)	(-2.73)	(-2.05)	(-2.16)	(-3.38)	(-1.84)
S&P 600	1996–2008 (1)		2.93		1.48		3.49
			(7.99)		(4.23)		(14.83)
	2009–2019 (2)		5.83		5.58		4.22
			(5.19)		(3.59)		(22.08)
	(2) - (1)		2.90***		4.10***		0.73***
			(2.63)		(2.75)		(4.49)

Panel B: Deletions							
		Call Options		Put Options		Stock	
		Transfer	Pure	Transfer	Pure	Transfer	Pure
S&P 400	1996–2008 (1)	0.40	0.47	0.70	2.23	1.26	3.71
		(0.78)	(0.81)	(1.00)	(1.22)	(3.90)	(3.87)
	2009–2019 (2)	0.97	-0.20	0.40	0.61	0.99	1.39
		(3.06)	(-0.78)	(1.01)	(1.11)	(6.18)	(2.13)
	(2) - (1)	0.57	-0.67	-0.30	-1.62	-0.27	-2.33**
		(1.05)	(-1.15)	(-0.42)	(-0.93)	(-0.75)	(-2.00)
S&P 500	1996–2008 (1)	2.03	0.51	3.55	1.43	1.92	6.34
		(2.03)	(1.43)	(2.46)	(2.35)	(6.66)	(3.75)
	2009–2019 (2)	0.97	1.54	0.25	2.43	0.80	4.03
		(1.92)	(0.99)	(1.27)	(0.96)	(7.08)	(1.42)
	(2) - (1)	-1.06	1.03	-3.29**	0.99	-1.12***	-2.31
		(-1.01)	(0.65)	(-2.47)	(0.38)	(-3.61)	(-0.70)
S&P 600	1996–2008 (1)		-0.32		1.12		6.62
			(-3.26)		(1.02)		(6.54)
	2009–2019 (2)		0.53		3.78		2.32
			(1.70)		(1.58)		(8.57)
	(2) - (1)		0.86		2.65		-4.30
			(2.82)		(1.11)		(-4.11)

Table 3.10: Incremental algorithmic trading around S&P index announcements

This table presents the median incremental change in algorithmic trading measures around the AD of S&P index additions and deletions. The Base Sample used runs from January 3, 2012, through December 31, 2019. The SEC MIDAS dataset is used, and I present the trade-to-order ratio and cancels-to-trade ratio. Lower values of trade-to-order are typical of higher algorithmic trading behavior, whereas higher values of cancels-to-trade are typical of higher algorithmic trading behavior. Stocks are classified as those that enter or leave the S&P 1500 (new addition or full deletion) or stocks that transfer between indexes (via promotion or demotion). I define three periods: pre-event (event time $t = -30$ to $t = -1$), event day, and post-event (event time $t = 1$ to $t = 30$). Incremental trading values are reported in percentage points. t-statistics are not reported, as most are statistically significant at the 1% or lower level.

Panel A: Additions							
		Trade-to-order			Cancels-to-trade		
	Event Range	S&P 400	S&P 500	S&P 600	S&P 400	S&P 500	S&P 600
Transfer	$[\tau_2, \tau_3]$	-5.23	0.49		-5.41	-8.91	
	τ_4	78.17	110.58		-45.74	-51.84	
	(τ_4, τ_5)	33.35	53.27		-29.51	-38.10	
	$[\tau_5, \tau_6]$	5.89	10.79		-15.54	-14.25	
Pure	$[\tau_2, \tau_3]$	-3.80	0.13	-4.31	-3.67	-6.27	-6.92
	τ_4	108.59	152.89	153.03	-48.93	-56.31	-63.23
	(τ_4, τ_5)	28.64	85.71	69.32	-31.64	-45.79	-43.48
	$[\tau_5, \tau_6]$	1.60	7.17	8.32	-11.73	-8.99	-18.08
Panel B: Deletions							
		Trade-to-order			Cancels-to-trade		
		S&P 400	S&P 500	S&P 600	S&P 400	S&P 500	S&P 600
Transfer	$[\tau_2, \tau_3]$	5.48	6.94		-14.00	-11.67	
	τ_4	37.42	42.29		-32.92	-30.25	
	(τ_4, τ_5)	13.38	26.24		-22.32	-23.11	
	$[\tau_5, \tau_6]$	8.59	10.59		-17.32	-10.74	
Pure	$[\tau_2, \tau_3]$	-2.15	-23.36	7.48	-7.50	15.30	-20.18
	τ_4	34.05	70.91	63.24	-20.94	-29.53	-45.95
	(τ_4, τ_5)	18.85	58.95	39.08	-25.92	-29.04	-33.21
	$[\tau_5, \tau_6]$	-5.00	10.15	17.14	-11.06	10.54	-24.50

Chapter 4

Less is more? Biases and overfitting in machine learning return predictions

4.1 Introduction

Machine learning models have been successfully employed to cross-sectionally predict stock returns using lagged stock characteristics as inputs. In this chapter I show that training market capitalization group–specific machine learning models produces superior stock-level return predictions and long–short portfolios. This result challenges the generally held belief that more training data lead to superior machine learning models for return prediction. Instead, machine learning models trained on the full cross-section of raw stock excess returns overfits to small stocks. This overfitting subsequently produces return predictions that produce inferior VW long–short portfolios. Applying appropriate target regularization, such as subtracting the cross-sectional median return within size groups, achieves similar performance improvements as training group–specific models without the added computational cost of training additional models. These results highlight the importance of the careful, guided application of machine learning models in the asset pricing setting.

Across a range of statistical models, I independently train three models on three non-overlapping groups of stocks: large-, mid-, and small-capitalization. Using each model, out-of-sample (OOS) return predictions are made and merged back into a full panel. For long–short zero-cost portfolios formed from an ensemble of sorted machine learning return predictions, I find an increase in the VW long–short portfolio Sharpe

ratio from 1.25 (full-trained model) to 1.52 (group–specific model) and an average increase in the cross-sectional information coefficient (IC) of 1.34% (from 6.08% to 7.42%). These increases are statistically significant and remain after controlling for known asset pricing models, such as the FF5 model (Fama and French, 2015).

To explore whether the outperformance of group–specific models is specific to the U.S. CRSP data setting used or is a more general feature of the application of machine learning in asset pricing, I conduct simulations using a data generating process (DGP) that models conditional relations between stock characteristics and stock returns. A characteristic that does not directly predict future returns but affects the level at which other characteristics predict returns is introduced into the DGP. I examine the effect of changing the level of volatility within different sub-samples of the DGP, which mirrors the higher volatility of small stocks compared with larger stocks in the real world. These simulations help shed light on the properties of the cross-sectional asset pricing characteristics that influence the degree to which machine learning models can predict returns across the entire cross-section of stocks. In the simulation setting, the group–specific approach to training machine learning models also outperforms the models trained on all data. Thus, the observed outperformance of group–specific models in the empirical setting is not only a feature of the U.S. CRSP data but also a feature of the modeling decisions used to design and estimate the machine learning models.

Finance literature has predominantly focused on the application of machine learning models for stock return prediction and portfolio allocation decisions (Moritz and Zimmermann, 2016; Heaton, Polson and Witte, 2017; Feng, He and Polson, 2018; Feng, Polson and Xu, 2018; Rasekhschaffe and Jones, 2019; Kelly, Pruitt and Su, 2019; Freyberger, Neuhierl and Weber, 2020; Gu, Kelly and Xiu, 2020; Kozak, Nagel and Santosh, 2020; Chen, Pelger and Zhu, 2023; Harvey and Liu, 2021; Leung, Lohre, Mischlich, Shea and Stroh, 2021; Azevedo and Hoegner, 2023; Avramov, Cheng and Metzker, 2022). Little theoretical or empirical work has explored the numerous machine learning design choices that must be made and how these choices interact with the asset pricing problem. Instead, the literature suffers from a dispersion and lack of consistency in modeling choices. In addition to the hyperparameters selected for model tuning, such as optimizer learning rates and regularization penalties, modeling decisions also include how to regularize stock characteristics and returns, which activation functions to use in neural network hidden layers, and even what universe of stocks to train on. The lack of consistency in modeling choices across the literature slows the advancement of machine learning in finance, as it obscures comparative analysis and the interpretation of what drives differences in results. I contribute to this stream of literature by exploring the impact of various modeling

decisions on the subsequent stock return predictions and long–short portfolios, and highlighting the significant dispersion in achievable portfolio results.

I run a series of machine learning experiments on a three-layer neural network; in each experiment, a single change to a model design choice is made. Machine learning design decisions are categorized into choices around the input features, the neural network architectures, and the target variable. The results show that basic design choices, such as batch normalization, can produce a significant variation in the performance of machine learning return prediction models. I find that the design of the target variable has the most significant impact on model performance. Appropriate regularization of the target variable generally results in superior model performance when using sorted return predictions to form long–short portfolios.

A lack of regularization of the target variable used—total excess return—drives the outperformance of group–specific machine learning models. The distributions of small-, micro-, and nano-stocks' excess returns exhibit higher volatility, skewness, and kurtosis, relative to large- and mega-stocks, which can bias the machine learning training process, effectively overfitting to predict the returns of these smaller stocks more accurately. Machine learning models aim to minimize prediction error under some loss function. Owing to the larger magnitude and dispersion of excess returns in smaller stocks, a machine learning model trained on a broader universe of stocks is more likely to overfit to smaller stocks. The smallest training loss can be achieved by more accurately predicting small stocks. However, this produces a predictive model that does not generalize to larger stocks, resulting in poorer performance, especially when forming VW portfolios. By separately training on three groups of stocks sorted by market capitalization, machine learning models overfit less to certain inherent biases, allowing to discover a stronger predictive model within each group–specific sample. Alternatively, adjusting the target total return variable through approaches such as subtracting the median excess return within size groups or rank normalizing excess returns achieves comparable performance to training within sub-samples but at a lower computational cost.

Numerous researchers have applied machine learning techniques for stock return prediction across various markets, including both developed and emerging markets (Choi et al., 2019; Tobek and Hronec, 2021; Chen et al., 2023; Azevedo et al., 2023; Cakici et al., 2023; Hanauer and Kalsbach, 2023) and specific countries (Drobetz and Otto, 2021; Leippold, Wang and Zhou, 2022; Rubesam, 2022; Lalwani and Meshram, 2022; Liu, Tao, Tse and Wang, 2022). These studies generally find that the results in the U.S. sample hold across other universes and regions, highlighting the robustness and power of machine learning models for return prediction. However,

despite these advances and strong results, relatively little work has explored the impact of machine learning methodological choices on the stock return prediction problem. By highlighting the dispersion of stock-level return prediction outcomes and portfolio outcomes that arise under different model design decisions, the variety of choices that must be made when implementing a machine learning model for stock return prediction and the impact each decision can have on the results is emphasized. My results suggest that asset pricing literature would likely benefit from a benchmarking approach to the general return prediction problem. This allows for a fairer and more consistent comparison of results across different research efforts.

Several recent studies explore different design aspects of machine learning models in finance, including the choice of performance measures used in loss functions (Dessain, 2022), the use of an online early stopping algorithm (Wong, Chan, Azizi and Xu, 2022), the effect of changing target excess return variables (Li, Simon and Turkington, 2022), and changing the prediction horizon (Blitz, Hanauer, Hoogteijling and Howard, 2023). Blitz, Hoogteijling, Lohre and Messow (2023) take a practitioner's perspective and demonstrate how methodological design choices impact the predictive outcomes of machine learning models. This study expands the dimension of decisions explored and demonstrates the significant dispersion in machine learning model performance.

My findings diverge from those of Chen et al. (2023), who find that their generative adversarial network (GAN) can efficiently capture the structure of large stocks when trained on a full cross-section of stocks. However, the model of Chen et al. (2023) is more complex than the neural networks used in seminal papers such as Gu et al. (2020) and may result in less overfitting. In addition, Chen et al. (2023) only train on stocks with all available characteristics, leading to a more homogeneous distribution that may represent the broader market. Training on stocks with all available characteristics acts as a form of regularization, as smaller stocks tend to have poorer coverage when computing various characteristics.

This chapter also contributes to the literature on group–specific modeling and conditional relations between asset pricing characteristics. Piotroski (2000) and Beneish, Lee and Tarpley (2001) demonstrate the advantages of using contextual analysis to study stock returns and predict extreme events. Sorensen, Hua and Qian (2005) argue that traditional asset pricing characteristics, such as the price-to-earnings ratio, can have varying efficacy for predicting stock returns depending on the risk context. Blitz and Hanauer (2020) show how small stock exposure can act as a catalyst for other factors and increase returns associated with investing in these portfolios. More recently, Cong, Feng, He and He (2022)

and Cong, Feng, He and Li (2022) have explored how machine learning models can uncover group–specific factor models that outperform when using tree-based models to split the cross-section of stocks. Qian and Su (2016) also demonstrate the power of group–specific modeling when using applying group fused Lasso to panel data models. I show how using market capitalization to create group–specific models can uncover interesting empirical results that challenge previously accepted results. Whereas the asset pricing literature typically searches for a single parsimonious model capable of explaining the cross-section of returns for all assets, for practitioners interested in forming investable portfolios with the highest OOS performance, a model that incorporates contextual analysis can be valuable. The properties of machine learning algorithms can allow for a richer and more flexible contextual asset pricing framework than standard econometric tools.

The results of this chapter have implications for academics and practitioners using machine learning models in finance, beyond stock return prediction. First, it highlights the biases and overfitting inherent in cross-sectional machine learning return predictions, confirming results from other research that group–specific models may deliver superior OOS performance owing to a lack of proper regularization. Second, it demonstrates the significance of various design choices in machine learning models and their impact on prediction outcomes, emphasizing the importance of data selection and prediction target design. Third, it provides valuable insights into the role of group–specific modeling and contextual analysis in understanding the cross-sectional relation between asset pricing characteristics and stock returns. Last, this study offers practical implications for academics and practitioners by emphasizing the importance of understanding the problem context and the impact of each step in the model design process, particularly when dealing with non-homogeneous return distributions. These contributions underscore the need for careful guidance when applying machine learning models in finance to avoid pitfalls and biases.

The rest of this chapter is structured as follows. Section 4.2 explains the data and method for estimating the machine learning models. Section 4.3 presents the empirical results from training group–specific machine learning models. Section 4.4 forms hypotheses to explain the results from Section 4.3 and presents the results from simulating different panels of factor data. Section 4.5 explores the impact of three dimensions of machine learning model design choices on the performance of long–short portfolios. Section 4.6 then concludes the chapter.

4.2 Machine learning setup and data

In designing and estimating machine learning models, I follow the general empirical setup of Gu et al. (2020). I use Chen and Zimmermann's (2022) Open Source Asset Pricing (OSAP) database for monthly stock-level characteristics, and I do not include any macroeconomic covariates in the study. In addition, I focus on group–specific machine learning models, where I separately train machine learning models for different size-groups of stocks.

4.2.1 Prediction problem

The general return prediction problem involves identifying a functional form $g(\cdot)$ of $\mathbb{E}(r_{i,t+1})$ that maps a set of m predictor variables for n stocks from $\mathbb{R}^{n\times m} \mapsto \mathbb{R}^n$ such that the maximum OOS predictive power relative to $r_{i,t+1}$ is obtained. I use a general additive prediction error model to describe this mapping between a stock's excess return and the set of predictor variables:

$$r_{i,t+1} = \mathbb{E}(r_{i,t+1}) + \epsilon_{i,t+1}, \tag{4.1}$$

where $r_{i,t+1}$ is the excess return of stock i in month $t+1$. I assume that the conditional expectation of $r_{i,t+1}$ is determined by some time-invariant function $g(\cdot)$ that takes a set of m predictor variables as input at time t:

$$\mathbb{E}(r_{i,t+1}) = g(z_{i,t}), \tag{4.2}$$

where $z_{i,t}$ is an m-dimensional vector of predictor variables, stocks are indexed by $i = 1, \ldots, N_t$ and months by $t = 1, ..., T$. I aim to find a functional representation of $g(\cdot)$ which maximizes OOS predictive power under some statistical model.

4.2.2 U.S. equities sample

U.S. equities data are from CRSP, and comprise all stocks listed on the NYSE, NASDAQ, and AMEX exchanges from March 1957 to December 2021. Excess returns are calculated as the one-month-forward stock returns over Treasury bill rates. I use the March 2022 version of the OSAP database of Chen and Zimmermann (2022) to source stock-level characteristics. This database replicates a large sample of cross-sectional stock return predictors from the literature. From this database, I use the 206 predictive factors and additionally specify Short-term Reversal as the prior one-month return, Size as the natural logarithm of market equity multiplied by price ($ME \times PRC$) from CRSP, and Price as the CRSP

PRC field. In addition to stock-level characteristics, 74 industry dummies based on the two-digit truncated Standard Industrial Classification (SIC) codes are included. The OSAP characteristics and SIC sector dummies are merged to produce a full panel of stock-month characteristics, denoted as Z. Non-categorical features are cross-sectionally rank normalized between -1 and $+1$, and the monthly cross-sectional median rank is used to fill missing data. For each month t in the sample, the vector of stock i's characteristics, $z_{i,t}$, is obtained from the corresponding month in the full panel Z.

In addition to the full panel of data, three separate groups are formed by splitting Z based on the Size variable. These panels are labeled as small, mid, and large, respectively. In each month t, a stock is allocated to one of the three groups based on its rank market capitalization at time t: bottom 30% (small), middle 40% (mid), and top 30% (large).[8] To maintain three reasonably balanced panels, I do not use NYSE breakpoints or any other capitalization split. If the sample were split based on raw market capitalization rather than rank, there would be substantially smaller panels for large- and mid-capitalization stocks, and the desired results would be less comparable.

4.2.3 Model frameworks

Several statistical models are trained to predict cross-sectional excess returns: linear models, regularized linear models, tree-based models, deep neural networks, and model ensembles. Specifically, I select an ordinary least squares (OLS) model, a regularized elastic net (ENET), a random forest (RF), gradient boosted regression trees (GBRT), and deep neural networks with 1–5 hidden layers (NN1-5). In addition, two ensemble models are formed. The first ensemble is the average stock-level prediction across all models (ENS). The second ensemble is the average stock-level prediction across the five deep neural network models (ENSNN).

The standard machine learning approach (Gu et al., 2020) from the literature is used to train the models, excluding OLS. The first OOS prediction is made for January 1987. The first 18 years of data (from March 1957 to December 1974) are used as only training data, and the next 12 years of data (from January 1975 to November 1986) are used as validation data for hyperparameter tuning. Predictions in the OOS test set are made from January 1987 to December 1987. A one-month gap is used between the end of the validation set and the start of the test set. Models are annually fit at the end of December, where hyperparameter tuning is performed, and OOS predictions for the following 12 months are made using the best-performing

[8] I find similar results using NYSE cutoff points at 30%/40%/30%.

model. After training the model and making OOS predictions, one year is added to the training sample (i.e., an expanding window). The validation sample is rolled forward by one year, including the most recent year of data and dropping the oldest year of data (i.e., a rolling window). For the neural network models, owing to the inherent randomness in the training algorithms, an ensemble approach is used, where the same model is trained 10 times with a different random seed. The final prediction used is the average of all 10 predictions. Section 4.3.3 explores the effect of model averaging on neural network prediction results.

For the OLS model, as there are no hyperparameters to tune, the validation set is included in the train set. The full set of hyperparameters used for each model can be found in Appendix 4.6.

4.2.4 Cross-sectional evaluation

I follow existing literature and introduce alternative evaluation measures to evaluate the OOS excess return predictions. The OOS pooled R^2, where the denominator is not demeaned, denoted as R^2_{OOS}, is calculated as:

$$R^2_{OOS} = 1 - \frac{\sum_{(i,t)\in OOS} \left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in OOS} r^2_{i,t+1}}, \tag{4.3}$$

where OOS refers to the testing sample period in which the data never enter the machine learning model for training or validation and $\hat{r}_{i,t+1}$ is the model prediction at t of the $t+1$ excess return for stock i. I also calculate R^2 where the denominator is demeaned with the monthly cross-sectional average excess return, denoted as R^2_{CS}:

$$R^2_{CS} = 1 - \frac{\sum_{(i,t)\in OOS} \left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2}{\sum_{(i,t)\in OOS} \left(r_{i,t+1} - \bar{r}_{i,t+1}\right)^2}. \tag{4.4}$$

In agreement with Wong et al. (2022), pooled OOS measures such as R^2_{OOS} and R^2_{CS} place higher importance on periods with higher numbers of stocks and are sensitive to outliers. In the context of forming long–short portfolios based on excess return predictions, the choice of R^2 as an evaluation metric does not fully align with the desired portfolio outcomes. As an alternative measure, I use the IC introduced by Ambachtsheer (1974) and further developed by Grinold (1989), which is the time-series average of the cross-sectional Pearson's correlation between excess returns and predictions:

$$IC = \rho\left(r_{i,t+1}, \hat{r}_{i,t+1}\right), \tag{4.5}$$

where ρ is Pearson's correlation. I note that the IC can be interpreted as the square root of the regression R^2, and thus out-of-sample we expect directionally similar results.

To test for differences in OOS predictive accuracy between two models, the Diebold and Mariano (1995) test for pairwise differences in prediction errors is applied. This specification compares the annual cross-sectional average of error in the testing sample predictions in each model rather than comparing the stock-level prediction errors. Estimating this model involves defining the test statistic DM_{12} as:

$$DM_{12} = \frac{\bar{d_{12}}}{\hat{\sigma}\left(d_{12}\right)}. \tag{4.6}$$

d_{12} is defined as:

$$d_{12,t+1} = \frac{1}{n_{3,t+1}} \sum_{i=1}^{n_3} \left(\left(r_{i,t+1} - \hat{r}_{i,t+1}^{(1)}\right)^2 - \left(r_{i,t+1} - \hat{r}_{i,t+1}^{(2)}\right)^2 \right), \tag{4.7}$$

where $\hat{r}_{i,t+1}^{(m)}$ is the excess return prediction from model $m \in (1, 2)$ for stock i at time $t+1$ and $n_{3,t+1}$ is the number of stocks in year $t+1$ (i.e., the testing sample for the model trained until December of year t). $\bar{d_{12}}$ and $\hat{\sigma}(d_{12})$ are the mean and Newey-West standard error of $d_{12,t}$ over the testing sample, respectively.

4.2.5 Portfolio evaluation

Standard portfolio statistics in the OOS period are calculated to evaluate portfolio performance. In all cases, I used the top-minus-bottom portfolio, defined as the difference in returns of decile portfolio ten and decile portfolio one, based on the sorted predicted returns at time t. First, the annualized Sharpe ratio is defined as:

$$SR = \frac{12 \times \mathbb{E}\left[P_t\right]}{\sqrt{12 \times Var\left[P_t\right]}}, \tag{4.8}$$

where P_t is the time series of monthly top-minus-bottom portfolio returns. Second, two drawdown measures are calculated. The first is the maximum portfolio drawdown, defined as:

$$MaxDD = \max_{0 \leq t_1 \leq t_2 \leq T} (Y_{t_1} - Y_{t_2}), \tag{4.9}$$

where Y_t is the cumulative log return from portfolio inception to time t. The second drawdown measure is the worst one-month portfolio drawdown: the largest negative portfolio return experienced in any one month in the OOS period. Next, the average monthly one-way portfolio turnover is calculated as:

$$Turnover = \frac{1}{T} \sum_{t=1}^{T} \left(\sum_{j} \left| w_{i,t+1} - \frac{w_{i,t}\,(1 + r_{i,t+1})}{1 + \sum_j w_{j,t} r_{j,t+1}} \right| \right), \tag{4.10}$$

where $w_{i,t}$ is the weight of stock i in a given portfolio at time t. For the top-minus-bottom portfolios, the maximum monthly one-way portfolio turnover is 200%, which corresponds to a complete replacement of all stocks in both the top and bottom portfolios.

Finally, one-way break-even transaction costs are calculated under two models: the FF6 model (Fama and French, 2015) and the HXZ model of Hou, Xue and Zhang (2015). Break-even transaction costs are defined as the average trading costs at which the alpha under each of these models becomes zero, where alpha is estimated using time-series regressions of the top-minus-bottom portfolio excess returns on the factor returns of the FF6 and HXZ models.

4.3 Group–specific models

The primary approach used in the extant literature for training machine learning models for cross-sectional stock return prediction uses the entire CRSP universe or some other pre-defined universe, such as MSCI World. This section presents the results from training group–specific machine learning models derived from the CRSP universe. An identical approach for model training and hyperparameter tuning is followed for the three size–specific groups. Predictions for the OOS period are made by each model and are then concatenated into a full cross-section of return predictions, resulting in a consistent set of predictions to the model trained on the full CRSP universe and allowing for a fair comparison between models.

Table 4.1: Sample statistics of monthly cross-sectional excess returns
This table presents the summary statistics of monthly cross-sectional excess returns for the in-sample training period (from March 31, 1957, to December 31, 1986) and for the OOS test period (from January 31, 1987, to December 31, 2021). The sample is split into three market capitalization groups: the largest 30% of stocks (Top 30%), the smallest 30% of stocks (Bottom 30%), and the middle 40% of stocks (Middle 40%). The table provides the time-series average of each group's excess returns and other relevant statistics.

Universe	Sample	NStocks	Mean	Std. Dev.	Min.	$q_{0.01}$	$q_{0.25}$	Median	$q_{0.75}$	$q_{0.99}$	Max.	Skew	Kurtosis
All	Test	7335	0.69	16.31	-90.48	-35.70	-5.81	-0.09	5.74	50.29	384.04	4.97	149.98
	Train	3522	0.77	11.81	-64.73	-23.96	-5.49	-0.22	5.67	38.06	149.09	1.85	23.99
Bottom 30%	Test	2358	0.83	22.04	-88.53	-43.39	-7.66	-0.61	6.16	72.45	371.50	4.80	92.57
	Train	1111	1.03	15.30	-61.39	-28.05	-7.16	-0.71	6.74	50.87	143.44	1.83	16.68
Middle 40%	Test	2628	0.56	14.10	-75.25	-32.99	-5.82	-0.13	5.82	43.67	159.45	1.73	33.52
	Train	1308	0.69	10.83	-46.10	-22.43	-5.54	-0.20	5.79	33.32	85.44	1.12	9.21
Top 30%	Test	2349	0.70	10.15	-56.91	-24.83	-4.48	0.42	5.50	29.67	85.48	0.66	11.20
	Train	1103	0.59	7.74	-31.86	-16.99	-4.04	0.21	4.77	22.04	49.58	0.62	5.21

Table 4.1 shows the time-series averages of the target excess returns each month, split into test (from January 1987 to December 2021) and train (from March 1957 to December 1986) sets and the three size groups. We observe many standard empirical results around smaller stocks, such as higher volatility, average excess returns, kurtosis, and skewness. These stylized facts of smaller stocks influence the statistical properties of the full sample. The skewness and kurtosis of the universe containing all stocks (denoted as All in Table 4.1) are higher than those of the three other size–specific universes. This demonstrates the effect of small stocks on the overall distribution of excess returns used to train the prediction models. The distributional characteristics of the excess returns can produce unintended biases when fitting machine learning models. For example, if the average return of some group of stocks (in the training data) is 2% higher than that of all other groups, under a mean squared error (MSE) loss function, a machine learning model will be rewarded more by minimizing the prediction error for these stocks, resulting in an inherent bias toward more accurately predicting the returns of stocks with a wider range of historical returns. The training of group–specific models aim to predict a more homogeneous distribution of excess returns, potentially reducing biases and overfitting toward the size–specific groups. This training approach of group–specific models act as a form of implicit target variable regularization.

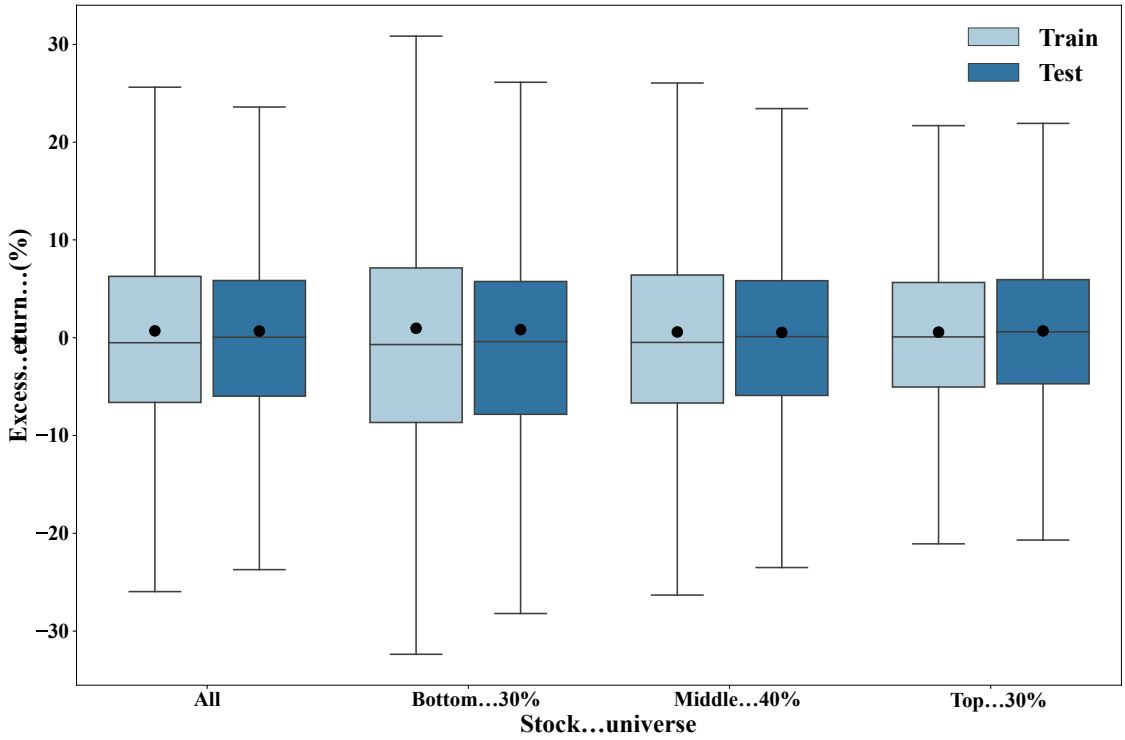

Figure 4.1: Excess return distributions across different size groups
This figure presents the box plots of monthly excess return distributions across different stock universes. The All universe contains all stocks, whereas the Bottom 30% contains the smallest 30% of stocks by market capitalization, the Middle 40% contains the middle 40% of stocks by market capitalization, and the Top 30% contains the largest 30% of stocks by market capitalization. Each box plot represents the distribution of monthly excess returns within each stock universe. The boxes represent the interquartile range (IQR), with the lower and upper boundaries of the box representing the 25th and 75th percentile of the distribution, respectively. The horizontal line within each box represents the median excess return. The whiskers extend to the minimum and maximum values within 1.5 times of the IQR. The solid black dot represents the mean excess return in each stock universe.

Figure 4.1 presents a box plot of the excess return distribution across the size–specific groups and the train–test data splits. The dispersion of the return distributions, on average, decreases between the train and test sets. This change in dispersion highlights a challenge when predicting excess returns using machine learning models. The distributional properties of excess returns are non-stationary, and the use of an expanding window approach incorporates all information into the prediction objective. This non-stationarity can produce unexpected biases in machine learning model predictions. For example, when predicting stock returns in December 2020, the machine learning model is trained on all data between March 1957 and December 2010, where each observation is equally contributing to minimizing the loss function. Thus, if the target return variable is not normalized (a common practice in the extant

literature), the machine learning model may overfit toward predicting returns earlier in the sample owing to the greater dispersion in excess returns. This produces a strong in-sample fit but can result in poor OOS performance. One alternative to this approach is to apply a weighting scheme to the training data, where more recent observations are weighted more. Another approach is removing the cross-sectional average monthly return (as a proxy for the market return) from each stock's monthly return.

4.3.1 Cross-sectional predictability

Table 4.2 shows the stock-level prediction comparison for the nine statistical models and two ensembles when trained on the full cross-section of returns (Full) versus the group–specific models (Size). The predictions from the Full and Size models are compared with the average of each stock-level prediction in these two models, denoted as the Ensemble model. Panel A presents the results from applying pairwise Diebold-Mariano (DM) tests across these models.[9] For the neural network models, the group–specific models generally have positive and statistically significant improvements in stock-level excess return predictions over the full models. This effect increases when ensembling the Size and Full model predictions. This improvement holds not only when calculating evaluation metrics in the whole CRSP universe but also when calculating the metrics within each size–specific universe. For example, suppose the predictive efficacy is evaluated in the large stocks only universe (Top 30%). In that case, the model specifically trained on large stocks outperforms the model trained on all stocks. The outperformance of group–specific models is a striking result, counter-intuitive to the expected behavior of machine learning models. Typically, the more data available and more examples for the model to learn from, the higher the predictive efficacy of the model. These results show that training machine learning models on size–specific groups produce return predictions superior to those of models trained on the full cross-section of stocks.

[9]Bonferroni corrections are not used, as I only compare pairwise models within each category of trained model: for example, the predictive efficacy of the NN5 model is not compared with that of the OLS model.

Table 4.2: Comparison of out-of-sample return predictions using Diebold-Mariano tests and information coefficient differences

This table compares OOS stock-level return predictions under different model training approaches using pairwise DM test statistics (Panel A) and IC differences (Panel B). I use three training regimes: training on the full cross-section (Full), training three group–specific models (Size), and averaging predictions from the Full and Size models (Ensemble). The sample is divided into two universes: All (combined sample) and Top 30%/Middle 40%/Bottom 30% (market capitalization splits). Positive values in Panel A indicate that the training approach in Model 2 performs better than in Model 1. Bold text indicates a statistically significant difference at the 5% level or lower.

Panel A: Diebold-Mariano test statistics

Universe	Model 1	Model 2	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
All	Full	Size	-1.34	**3.70**	-1.58	-0.15	**2.78**	**2.90**	**2.66**	**3.01**	1.69	0.49	**2.78**
	Full	Ensemble	-1.05	**4.80**	-0.34	**2.52**	**4.67**	**5.03**	**4.42**	**4.85**	**3.43**	**2.71**	**4.40**
	Size	Ensemble	1.44	-2.38	**2.57**	1.65	-0.12	-0.10	-0.40	-0.77	0.16	1.08	-0.81
Top 30%	Full	Size	-1.23	**3.03**	-1.84	**2.30**	**2.64**	**2.30**	**2.08**	1.95	1.35	0.28	1.93
	Full	Ensemble	-0.90	**3.70**	-0.95	**4.88**	**4.35**	**4.24**	**3.41**	**2.53**	1.49	**2.33**	**3.07**
	Size	Ensemble	1.34	-2.15	**2.30**	0.71	-0.51	-0.38	-0.53	-0.82	-0.92	0.90	-0.64
Middle 40%	Full	Size	-1.18	**2.92**	-1.86	0.51	**3.07**	**3.19**	**2.90**	**3.12**	**2.62**	0.71	**2.96**
	Full	Ensemble	-0.67	**3.66**	-0.85	**3.35**	**4.52**	**4.97**	**4.21**	**4.15**	**3.19**	**2.16**	**4.23**
	Size	Ensemble	1.34	-2.07	**2.67**	**2.10**	-0.96	-1.08	-1.07	-1.37	-1.52	0.79	-1.34
Bottom 30%	Full	Size	-1.47	**3.02**	-0.47	-0.62	1.88	**2.23**	1.89	**2.40**	0.68	0.28	2.17
	Full	Ensemble	-1.30	**4.24**	0.37	0.94	**3.85**	**3.64**	**3.70**	**4.29**	**2.05**	**2.41**	**3.87**
	Size	Ensemble	1.53	-1.62	1.24	1.28	0.60	-0.56	0.13	-0.32	1.16	1.11	-0.30

Panel B: Information coefficient differences

Universe	Model 1	Model 2	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
All	Full	Size	**1.37**	**2.11**	-0.07	0.12	**1.85**	**1.94**	**1.96**	**2.07**	**1.60**	**1.58**	**1.96**
	Full	Ensemble	**0.94**	**1.50**	0.56	0.66	**1.57**	**1.64**	**1.58**	**1.63**	**1.39**	**1.36**	**1.52**
	Size	Ensemble	-0.43	-0.61	0.63	0.54	-0.28	-0.30	-0.37	-0.44	-0.21	-0.22	**-0.44**
Top 30%	Full	Size	**1.81**	3.12	-0.06	0.37	**1.57**	1.28	1.47	**1.69**	1.16	**2.04**	**1.62**
	Full	Ensemble	**1.03**	**1.76**	0.32	0.18	**1.18**	**0.95**	**1.09**	**0.89**	**0.95**	**1.33**	**1.11**
	Size	Ensemble	-0.79	-1.36	0.38	-0.19	-0.39	-0.33	-0.38	-0.80	-0.21	-0.70	-0.50
Middle 40%	Full	Size	**1.73**	1.79	-0.56	1.20	**1.91**	**1.77**	**2.04**	**2.10**	**2.34**	**2.06**	**2.03**
	Full	Ensemble	**1.15**	**1.49**	0.24	0.92	**1.43**	**1.18**	**1.45**	**1.32**	**1.56**	**1.54**	**1.40**
	Size	Ensemble	-0.58	-0.29	0.79	-0.28	-0.46	-0.56	-0.56	-0.75	-0.74	-0.51	-0.59
Bottom 30%	Full	Size	**1.22**	1.19	1.13	-0.44	**1.63**	**2.00**	**1.90**	**2.10**	**1.37**	**1.20**	**1.91**
	Full	Ensemble	**0.89**	1.03	1.65	0.58	**1.51**	**1.50**	**1.63**	**1.70**	**1.17**	**1.19**	**1.54**
	Size	Ensemble	-0.33	-0.16	0.53	1.02	-0.11	-0.49	-0.27	-0.40	-0.20	0.00	-0.37

Panel B in Table 4.2 presents a similar comparison as Panel A but uses the IC instead. The null hypothesis that the difference in time-series ICs between Model 1 and Model 2 is zero is tested using yearly time series of ICs for each model. The same conclusion reached for the DM test statistics is reached using the IC. For neural network architectures, training on size–specific groups produce higher ICs. Table 4.2 also shows statistically significant improvements for the OLS model when trained in size–specific groups. For the tree-based models, the results are mixed. The regularized behavior already embedded into the architecture of tree-based models that helps to prevent overfitting could explain this result. However, this tree-based regularization appears to come at the cost of lower performance compared with the neural network models.

4.3.2 Portfolio performance

Long–short portfolios are formed by sorting stocks each month based on the excess return predictions and used to evaluate the portfolio performance of the prediction models. Decile portfolios are used, denoting portfolio ten as the long portfolio and one as the short portfolio. Long–short portfolios are formed using predictions at the end of month t, where \$1 is invested into the long portfolio and \$1 is invested into the short portfolio. These portfolios are held for one month and earn the VW or EW return.

Table 4.3 presents the average VW portfolio statistics. I report the annualized portfolio return, annualized portfolio volatility, Sharpe ratio, maximum portfolio drawdown, maximum one-month portfolio loss, one-way portfolio turnover, FF6 alpha and associated t-statistic, and FF6 and HXZ break-even transaction costs. Each of these measures is reported for the nine predictive models and two ensembles, as well as for the models trained on the full cross-section (Full), group–specific models (Size), and an ensemble of the Full and Size models (Ensemble).

Table 4.3: Out-of-sample performance of group–specific machine learning portfolios

This table presents VW top-minus-bottom portfolio statistics for machine learning models trained using three approaches: Full, Size, and Ensemble. The Full model follows the standard approach using all available stocks, whereas the Size model trains three separate models for large, mid, and small stocks and then concatenates the predictions to form portfolios. The Ensemble model is the average of the return predictions from the Full model and Size model. The table presents the performance statistics of these portfolios, including the annualized mean, standard deviation, Sharpe ratios, maximum drawdown, maximum one-month loss, average monthly one-way turnover, annualized FF6 alpha and t-statistic, and break-even transaction costs under the FF6 and HXZ risk models. The OOS period is from January 1987 to December 2021.

Metric	Train Approach	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Portfolio return (ann. %)	Full	12.4	6.5	10.9	3.8	16.0	16.1	16.4	19.7	20.4	19.0	20.0
	Size	15.7	28.2	9.6	19.4	27.5	30.7	27.1	29.6	23.5	30.1	31.8
	Ensemble	17.2	20.6	15.3	19.4	23.4	27.2	27.6	29.1	27.8	26.3	28.8
Volatility (ann. %)	Full	15.9	12.2	13.8	19.6	14.7	15.4	15.5	17.2	17.4	16.0	15.9
	Size	23.0	16.0	15.2	17.7	19.9	19.7	20.6	19.5	19.5	19.5	20.8
	Ensemble	20.8	16.1	16.0	17.7	17.2	19.4	19.4	18.8	17.7	18.3	19.0
Sharpe ratio	Full	0.78	0.53	0.79	0.19	1.08	1.04	1.06	1.14	1.17	1.19	1.25
	Size	0.68	1.77	0.63	1.1	1.38	1.55	1.32	1.52	1.21	1.54	1.52
	Ensemble	0.83	1.28	0.95	1.09	1.36	1.41	1.42	1.55	1.57	1.44	1.52
Max. drawdown (%)	Full	56.1	34.4	42.9	157.0	30.7	39.2	28.3	36.2	33.7	35.0	25.4
	Size	82.0	28.7	67.7	30.5	56.4	41.3	33.8	42.0	56.4	28.4	47.2
	Ensemble	95.6	35.8	70.1	29.1	28.1	36.5	29.5	41.0	32.5	42.3	34.2
Max. 1M loss (%)	Full	20.9	12.5	10.9	28.5	14.9	20.3	18.2	19.8	20.4	18.6	18.6
	Size	33.2	15.6	15.2	13.3	17.2	18.8	20.1	24.4	23.1	17.8	19.4
	Ensemble	32.3	13.9	14.6	12.8	17.4	19.6	18.2	19.0	16.1	16.5	19.0
Monthly one-way turnover (ann. %)	Full	124.8	118.4	117.3	156.8	144.2	146.7	147.8	149.6	150.2	150.9	149.9
	Size	120.9	153.8	122.2	155.6	143.9	143.3	142.1	141.8	138.5	148.4	142.6
	Ensemble	124.7	166.8	123.4	160.7	149.7	150.0	149.3	149.3	148.5	151.9	150.7
FF6 α (ann. %)	Full	10.9	5.2	10.9	4.1	15.3	13.4	15.4	18.7	18.0	17.2	18.8
	Size	13.6	27.8	7.3	18.0	25.9	29.1	24.1	27.0	21.5	28.2	29.8
	Ensemble	14.2	19.4	13.4	19.7	22.1	25.0	24.5	26.7	24.3	23.9	26.1
FF6 t-stat	Full	3.55	2.35	4.28	1.21	4.79	5.04	5.69	5.45	5.68	6.02	5.71
	Size	2.83	7.71	2.51	4.92	5.89	6.54	6.22	7.28	5.17	6.97	6.68
	Ensemble	2.99	6.38	4.57	5.41	6.85	6.73	6.67	6.72	7.05	6.59	6.77
FF6 break-even cost (bps)	Full	36.3	18.2	38.7	10.9	44.1	38.1	43.6	52.1	50.0	47.4	52.3
	Size	46.8	75.3	24.9	48.2	75.0	84.6	70.8	79.5	64.6	79.2	87.0
	Ensemble	47.6	48.6	45.1	51.0	61.5	69.4	68.4	74.5	68.2	65.6	72.1
HXZ break-even cost (bps)	Full	32.4	17.9	39.0	7.9	38.1	34.6	42.0	48.2	48.5	45.1	47.5
	Size	44.5	73.7	23.2	43.0	70.1	78.6	67.1	78.2	59.2	75.8	82.6
	Ensemble	48.8	47.1	45.1	46.3	56.6	64.3	66.7	72.3	66.9	61.2	70.3

Table 4.3 demonstrates consistently stronger portfolio characteristics when training on group–specific models. For the ensemble of all models, the average annualized portfolio return increases from 20.0% for the Full model to 31.8% for the Size model. An increase in portfolio volatility accompanies this, but this increase in risk is compensated for, as seen by the increase in Sharpe ratio from 1.25 to 1.52. Although the increase in portfolio volatility could be attributable to a greater presence of small stocks in the top and bottom portfolios, as the portfolios are VW, the overall impact of small stocks is reduced. Portfolio turnover does not significantly change but based on the increase in the FF6 break-even cost (increase from 32.3 to 54.1

bps) and HXZ break-even cost (increase from 27.9 to 40.0 bps), the increase in outperformance is not simply attributable to an increase in portfolio trading.[10] The analysis here is repeated within the three size universes. These results are reported in Appendix 4.6, and in general, the average long–short portfolio returns increase when using predictions from the model trained on the size–specific group (compared with predictions from the model trained on the full universe of stocks); this translates to higher Sharpe ratios and higher break-even transaction costs for the group–specific models.

Table 4.4: Sharpe ratio differences between machine learning training approaches

This table presents the OOS Sharpe ratio differences between the machine learning training approach specified in the Model column and the Full training approach using the Ledoit and Wolf (2008) Sharpe ratio test. Bold text indicates statistical significance at the 5% level.

Universe	Model	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
All	Size	-0.102	**1.250**	-0.116	**0.886**	**0.315**	**0.536**	0.301	**0.424**	0.019	**0.385**	0.281
	Ensemble	0.054	**0.769**	0.174	0.914	**0.304**	**0.371**	**0.401**	**0.404**	**0.401**	**0.245**	**0.271**
Top 30%	Size	-0.026	**0.806**	0.022	**0.536**	**0.259**	**0.509**	**0.420**	**0.476**	0.234	**0.398**	**0.381**
	Ensemble	0.052	**0.646**	0.207	**0.667**	**0.196**	**0.349**	**0.325**	**0.259**	**0.322**	**0.292**	**0.260**
Middle 40%	Size	-0.149	**0.400**	-0.453	0.302	0.088	0.004	0.119	0.095	-0.043	-0.141	0.116
	Ensemble	-0.029	**0.692**	-0.185	**0.743**	**0.215**	0.121	**0.164**	0.107	0.130	0.066	**0.141**
Bottom 30%	Size	-0.015	0.211	-0.920	0.178	0.032	-0.051	-0.151	-0.014	-0.162	0.100	-0.232
	Ensemble	0.005	**0.856**	-0.329	**0.247**	0.090	0.105	-0.071	**0.278**	0.015	0.238	-0.095

Although the increases in portfolio returns and Sharpe ratios are large, the statistical significance of the differences between these portfolio results cannot be gauged here. In Table 4.4, results for the Ledoit and Wolf (2008) test for differences in Sharpe ratios are reported. The Base model is set as the model trained on the full universe of stocks and then compared with the Size and Ensemble models. On average, the Sharpe ratio differences are positive and statistically significant when comparing neural networks trained on the full universe of stocks and neural networks trained in size–specific groups.

[10] In Table 4.13 in Appendix 4.6, the results also hold for EW portfolios.

Table 4.5: Factor loadings of machine learning portfolios
This table presents the results of regressing VW top-minus-bottom portfolio returns on the FF6 return series. Panel A presents the results for models trained on all stocks. Panel B presents the results for models trained in size–specific groups. The table reports the t-statistics for the regression intercept, and bold text indicates statistical significance at the 5% level.

Panel A: Full-trained model											
Factor	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Const.	0.009	0.004	0.007	0.004	0.010	0.011	0.01	0.014	0.016	0.013	0.013
	(4.50)	(2.06)	(3.65)	(1.39)	(4.73)	(4.89)	(5.02)	(5.70)	(6.63)	(5.53)	(5.83)
Mkt-RF	**-0.215**	0.034	0.036	**0.193**	-0.042	-0.074	-0.037	-0.097	-0.107	-0.047	-0.031
HML	0.175	**0.215**	**0.350**	-0.109	0.250	0.182	0.204	0.133	0.166	0.216	0.214
SMB	**-0.311**	0.108	**0.460**	**0.286**	0.132	0.015	0.104	0.030	0.090	0.024	0.059
UMD	**0.275**	**0.177**	0.071	-0.224	**0.305**	**0.250**	**0.350**	**0.495**	**0.414**	**0.207**	**0.348**
RMW	**0.585**	0.252	0.171	-0.158	**0.414**	**0.552**	**0.557**	0.179	0.055	**0.641**	**0.577**
CMA	**-0.260**	-0.246	0.142	**-0.525**	-0.068	-0.112	0.016	0.095	0.027	-0.346	0.023
R^2	43.2%	7.4%	16.1%	21.0%	18.4%	20.5%	25.2%	24.%	15.6%	18.6%	25.6%
Panel B: Size-trained model											
Factor	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Const.	0.007	0.020	0.008	0.016	0.019	0.022	0.017	0.021	0.014	0.020	0.021
	(3.27)	(9.42)	(3.57)	(4.64)	(7.82)	(8.12)	(8.10)	(8.55)	(6.52)	(7.50)	(9.21)
Mkt-RF	-0.063	-0.093	-0.125	0.030	-0.130	-0.112	-0.069	**-0.143**	-0.055	-0.065	-0.107
HML	**0.235**	**0.354**	0.071	0.046	**0.425**	**0.435**	**0.440**	**0.416**	**0.364**	0.206	**0.512**
SMB	**-0.602**	**0.268**	**0.588**	0.028	-0.212	-0.105	-0.168	-0.062	-0.030	0.077	-0.094
UMD	**0.393**	**0.432**	**-0.308**	-0.080	**0.265**	**0.349**	**0.254**	0.181	**0.260**	**0.298**	**0.290**
RMW	**1.375**	**0.367**	**0.352**	-0.025	**0.651**	**0.627**	**0.930**	**0.785**	**0.937**	**0.920**	**0.917**
CMA	-0.132	0.238	**0.343**	0.287	0.397	0.302	0.472	0.406	0.427	0.473	0.438
R^2	65.7%	33.0%	24.4%	0.8%	39.7%	35.6%	47.9%	40.1%	41.9%	35.2%	47.5%

By training on small stocks separately, larger predictions for these stocks can be made, and when these predictions are combined with the larger stock predictions, the portfolios may implicitly overweight small stocks. Table 4.5 shows the regression coefficients and intercept (alpha) obtained when regressing top-minus-bottom portfolio returns on the FF6 returns. The group–specific models do not have higher exposures to the Small-Minus-Big (SMB) factor than the full-trained models, suggesting that the size-trained models are not earning their increased return from taking greater exposure to a size premium. The group–specific model alphas for ENS and ENSNN have higher statistical significance and are larger in magnitude than the full model alpha. Despite the FF6 asset pricing model generally explaining more of the size-trained model returns (through the higher R^2), the alphas are larger in the size-trained models. The size-trained models obtain greater exposure to the FF6 asset pricing models whilst incorporating additional covariates that are not contained in the FF6 model.

The results found here present an anomaly. Common practice in machine learning literature suggests that the more data available to train a machine learning model, the higher the expected in-sample and OOS performance. I find the opposite.

Machine learning models trained to predict excess returns within size-groups outperform those trained to predict excess returns in the full cross-section. The rest of this chapter explores this anomaly in the context of machine learning models applied to the asset pricing setting.

4.3.3 Effect of model averaging

When training machine learning models (particularly neural networks), one problem is the instability of outcomes and inability to replicate results when training models. Instability emerges from the inherent randomness involved with training machine learning models and particularly the randomness associated with the chosen optimizer and its approach to minimizing the loss function. In previous literature (Gu et al., 2020; Wong et al., 2022), the authors train the same neural network model multiple times with a different random seed and then use the average prediction across each model. The number of models is arbitrary but is an important choice given the high time and computational costs associated with training machine learning models. The outperformance of group–specific machine learning models could be driven by the inherent randomness of the neural network training process. Thus, to test the robustness of this result, I run a quasi-Monte Carlo experiment.

Using the three-layer neural network architecture (NN3), the model is trained ten times for each of the four data samples (full, small, middle, and large). The uncertainty in performance associated with averaging across the different predictions is measured using a quasi-Monte Carlo experiment. In particular, the below procedure is followed:

1. Generate a random sequence of the 10 models by sampling without replacement,
2. Step through the sequence generated in (1) and take the average of all predictions at each step,
3. For each ensemble of predictions at each step, calculate the portfolio statistics,
4. Repeat steps (1)–(3) 100 times.

For example, suppose the random sequence generated in step (1) is (4, 5, 3, 10, 9, 2, 1, 8, 6, 7). I start with the fourth trained model and calculate the average top-minus-bottom portfolio return and average IC; this is then the first model. I then take both the fourth and fifth trained models, take the average prediction across both models, and estimate the two statistics; this is the second model. This

process is repeated until the final iteration that averages the predictions across all 10 models. This process simulates the effect of ensembling across all models, and the effect that starting with a specific model has on the overall outcomes. Figure 4.2 presents the results of this process for the NN3 model.

Figure 4.2 presents several different results. In Panel B as the number of models in the ensemble increases the average IC increases and the standard deviation of ICs decreases. These improvements demonstrate a common model ensembling result (Hansen and Salamon, 1990; Dietterich, 2000) and the effect of ensembling on predictive accuracy. This effect holds across all samples. The group–specific models always outperform the full sample models across the different number of models in the ensemble. Importantly, for the sample containing All stocks, even when accounting for the standard deviation in the ICs between the separate models, there is no overlap between the group–specific and full-trained models. The superior performance of the group–specific models is not a statistical fluke from how the machine learning models were trained but is a persistent feature of the models.

Panel A shows that the ensembling result holds for the top-minus-bottom VW portfolio returns. For small stocks, the impact of ensembling is greater than for all the other categories. The average top-minus-bottom return increases from 40% to 50% for the small stock predictions derived from the full-trained model and from 48% to over 60% for the small stock predictions derived from the model trained only on small stocks. This difference is likely a function of the higher level of volatility in smaller stocks. Thus, the increased predictive accuracy from ensembling across multiple predictions translates to higher OOS returns.

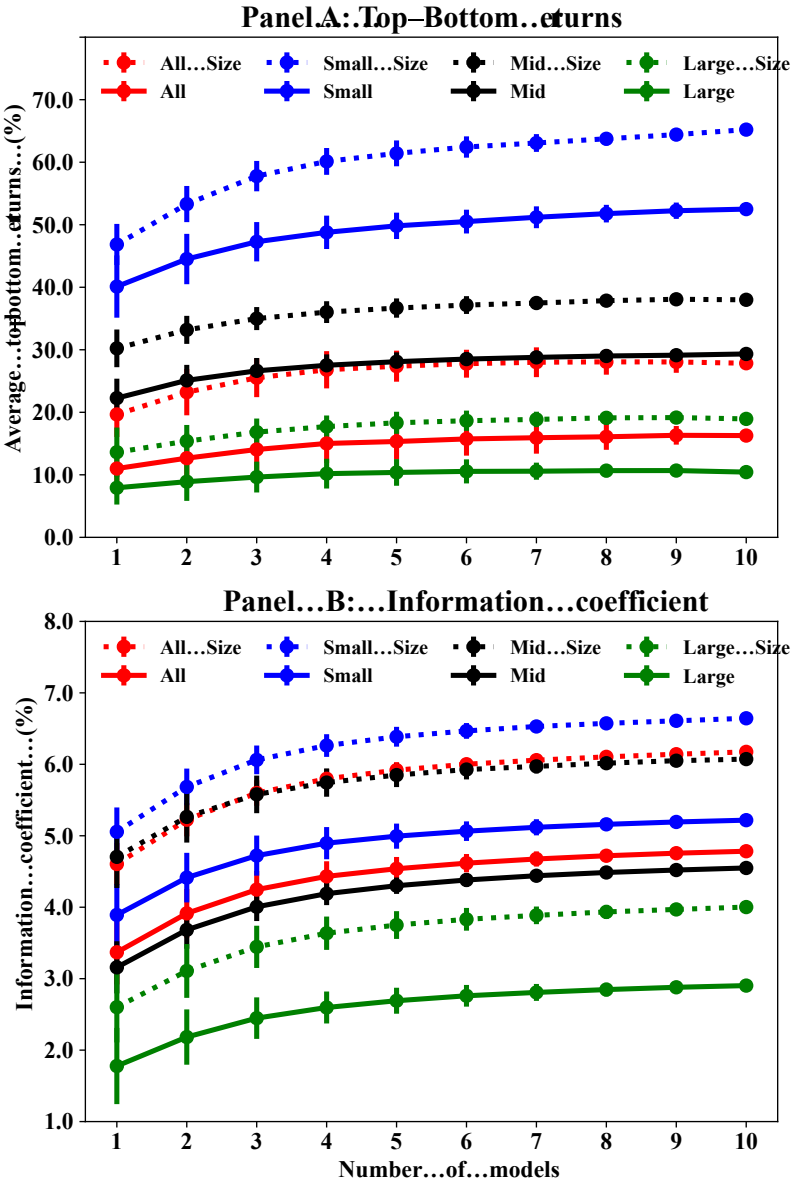

Figure 4.2: Effect of model averaging on neural networks
This figure presents the average annualized top-minus-bottom portfolio return (Panel A) and average IC (Panel B) for the NN3 model when ensembling identically trained models with different random seeds. The x-axis represents the number of models included in the ensemble. The solid lines present the results when training the NN3 model once using all stocks. The dashed lines present the results when training the NN3 model in three separate group–specific models. Ten models are trained using an identical approach but with different randomized seeds for each size group. A quasi-Monte Carlo approach is used to measure the effect of ensembling on the uncertainty of performance statistics. This approach is detailed in Section 4.3.3. This figure presents the average and standard errors of top-minus-bottom returns and ICs of the portfolios formed from these ensemble predictions.

4.3.4 Feature importance

We have seen the improvement in stock-level return predictions and the subsequent superior portfolio performances achieved when training group–specific machine learning models. Now I explore what stock covariates may be driving this improvement. To calculate feature importance, a standard partial dependence approach is used based on the change in predictive R^2 that occurs when setting all values of a predictor to zero, while all other predictor values remain unchanged. Feature importance is measured based on the change in R^2 when the predictor values are set to zero within the training sample. For each model, I calculate the change in R^2, normalize the value between zero and one, and take the average feature importance value across time.

Figure 4.3 presents the difference in feature importance between the group–specific models and the model trained on all data. In particular, I calculate the average time-series difference in feature importance for each variable, and in Figure 4.3, I report for each of the small (TSML), middle (TMID), and large (TLGE) models the 15 largest increases and decreases in feature importance. There is a marked change in the feature importance within group–specific models. One example is the SmileSlope variable, an options-derived variable. This variable has higher coverage for large stocks. Thus, when stocks that do not have coverage are removed from the training sample, the predictive efficacy of this feature, relative to other features, increases, and it increases in overall importance to the model.

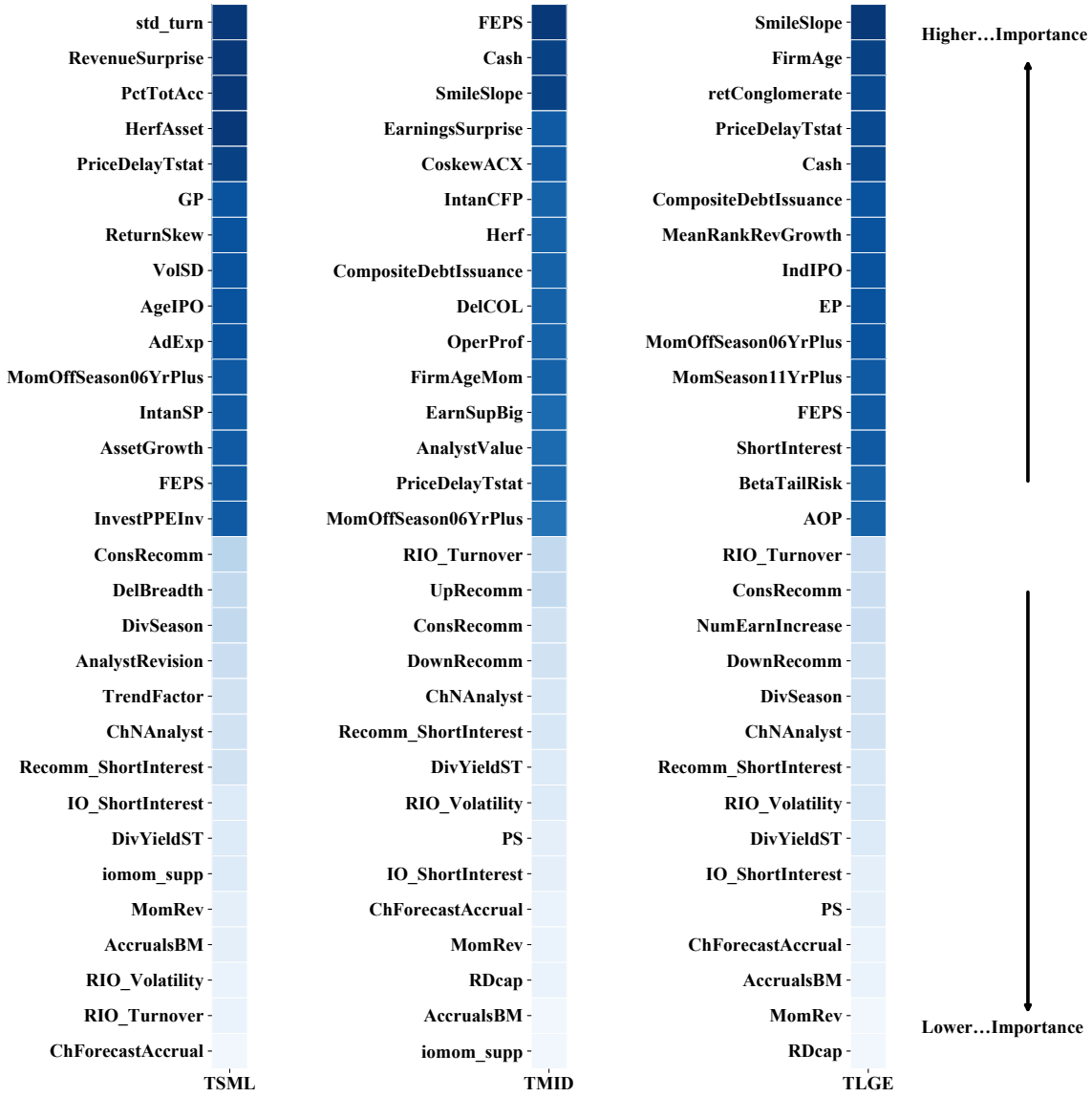

Figure 4.3: Feature importance when training in size categories
This figure presents the change in feature importance for group–specific neural network models compared with the feature importance of the model trained on all stocks. Higher importance indicates the features with the largest increase in average feature importance when compared with the full model. Lower importance indicates the features with the largest decrease in average feature importance. The results presented are the average across the five neural network models and through the OOS period from January 1987 to December 2021.

4.3.5 Covariate interactions

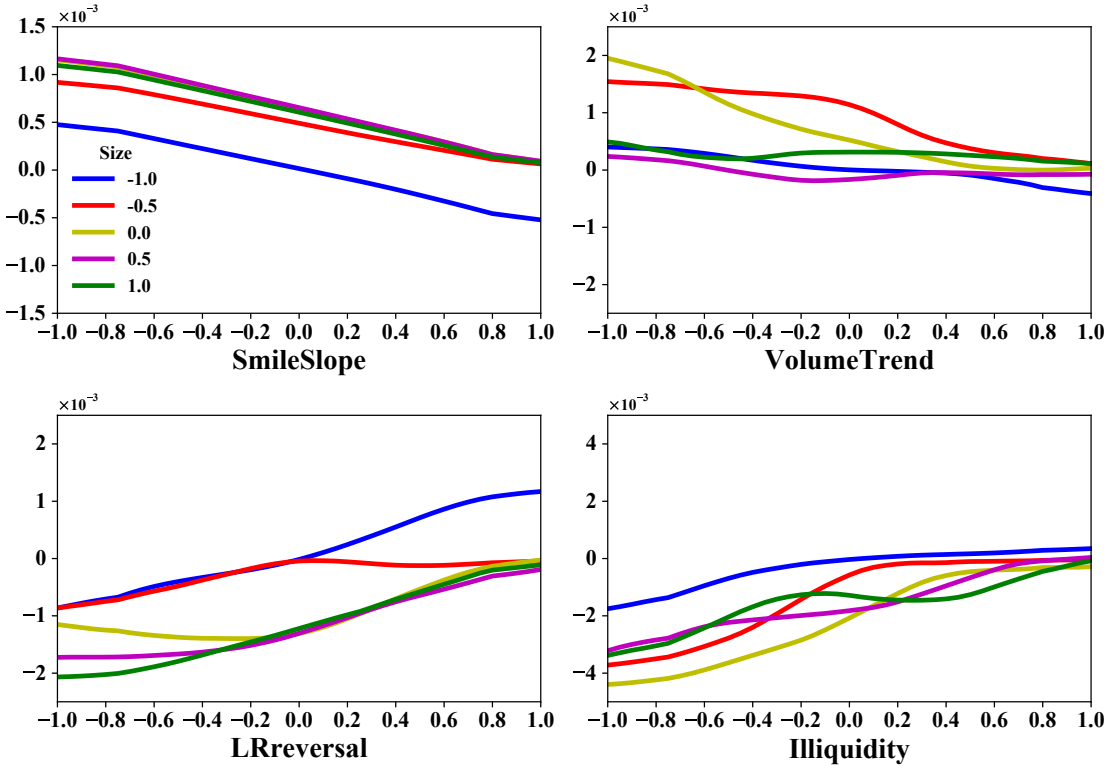

Figure 4.4: Size-conditional interaction between characteristics and model predictions
This figure illustrates the sensitivity of the NN3 model's excess return predictions to interactions between Size and four other covariates: SmileSlope, VolumeTrend, LRreversal, and Illiquidity. For each covariate, the value of Size is fixed, and the covariate value is iterated between -1 and $+1$ while holding all other covariates at their median value of zero. This figure then plots the average effect on return predictions over time on the y-axis. The x-axis represents the value of the covariate, whereas the y-axis represents the predicted excess return.

Having established that group–specific machine learning models use cross-sectional characteristics differently, I now look at the effect of covariate interactions within these models. I focus on the interactions with Size, given the choice to use this characteristic to train group–specific models. Figure 4.4 shows, for the model trained on the full dataset, the marginal effect of four features for five given levels of Size: SmileSlope, VolumeTrend, LRreversal, and Illiquidity. There is a degree of change for different values of the Size feature and the behavior of the primary feature. However, these interactions generally have similar impacts on the model's predictive output.

Figure 4.5 paints a different picture. It presents the marginal effect of the same four features taken from the independently trained models. We now see a significant

divergence between the impact of different features on model predictions across Size. In each of these selected examples, the impact of a feature on model predictions is the opposite between the large and small trained models. For SmileSlope, the coverage of this variable for stocks in the small sample is typically low; thus, the overall effect on the model should be quite limited, which is what we see. Contrasting this to Figure 4.4, the model has likely learned from the large stocks with coverage such that this variable is predictive of returns, and this variable is assumed equally predictive for small stocks. Similarly, for Illiquidity, the model trained only on large stocks has a flat effect on the predictive model for large stocks, whereas the effect is much more pronounced for small stocks.

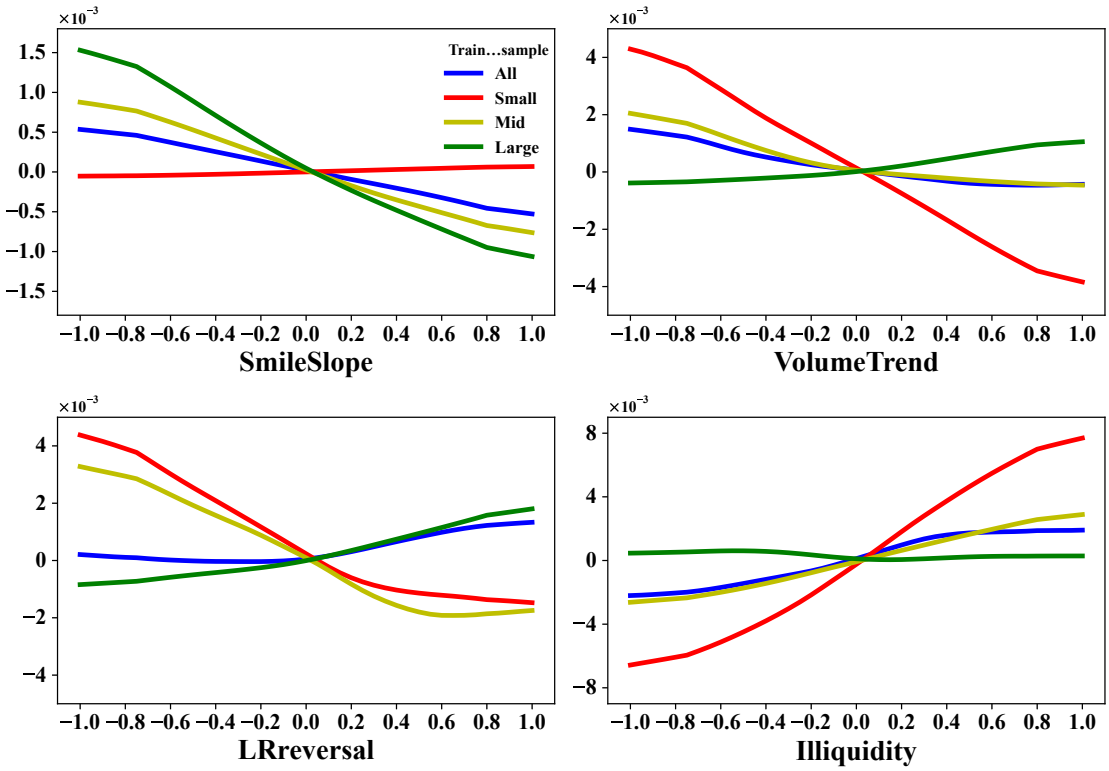

Figure 4.5: Marginal effect of covariates on excess return predictions across different training samples
This figure presents the sensitivity of the NN3 model's excess return predictions for four covariates under different training approaches. For each model, all other covariates are fixed at their median value of zero, and the model predictions are calculated for all values between -1 and $+1$ for the chosen covariates. The x-axis represents the value of the covariate, and the y-axis represents the predicted excess return. Different colors represent different training samples used to fit the NN3 model.

The data the model has been trained on can significantly influence the relation between an input feature and the model prediction. A machine learning model simply minimizes a given loss function. If certain characteristics in the training data

result in lower overall model losses, then this is what the machine learning model will learn to use to predict returns. However, the subsequent use and application of these predictions may not be fully aligned with the process used to obtain these predictions. In the case of forming long–short portfolios, this can then result in sub-optimal OOS performance.

4.4 Simulation study

The outperformance of group–specific machine learning models poses a challenge to the commonly held belief that more training data lead to superior performance of machine learning models. To assess whether this anomaly is primarily a feature of the U.S. CRSP data setting or a generalized result for machine learning models, I conduct a simulation study using group–specific dependencies between simulated input features (stock characteristics) and outputs (stock returns) and vary the levels of volatility and predictive efficacy within these groups. I follow the basic DGP setup from Gu et al. (2020) with augmentations that simulate a conditional dependence between covariates. Appendix 4.6 contains the full details of the simulation approach.

4.4.1 Hypothesis formation

Table 4.1 shows the more pronounced volatility and kurtosis of excess returns of the bottom 30% of stocks in the CRSP sample. The first hypothesis for why group–specific machine learning models outperform is that a machine learning model may overfit to specific groups of stocks (such as small- and micro-capitalization stocks) owing to the inherent extreme characteristics of the return distribution of these stocks. For example, a higher variance of excess returns combined with large absolute magnitudes of returns potentially provides more opportunity to improve the overall loss function used to train a machine learning model. This hypothesis posits that the machine learning model does not generalize as strongly to other groups of stocks (such as large- and mid-capitalization stocks), and this is then mitigated by training group–specific models. The predictive efficacy and volatility can be varied between groups within the training dataset to assess this hypothesis. If the model overfits to a specific group in the training dataset, we expect to see a variance in the predictive efficacy across the groups.

- Hypothesis 1: model overfits to a specific group of stocks.

The second hypothesis states that different groups of stocks have different return drivers that are difficult for machine learning algorithms to capture, even when

features that identify these groups (such as market capitalization) are in the input feature set. I assume a prior belief that the asset pricing characteristics that drive cross-sectional returns vary across size groups and can be introduced into the machine learning model design by training group–specific models. Under this hypothesis, group–specific models are more efficient at capturing this variation in cross-sectional return drivers, compared to models trained on the full dataset, and achieve superior OOS performance when using these predictions to form long–short portfolios. The levels of factor predictability in the group–specific DGPs are varied to study this hypothesis in the simulation.

- Hypothesis 2: stock groups have different drivers of returns that machine learning algorithms struggle to efficiently capture.

4.4.2 Simulation design

To simulate the effect of conditional relations on machine learning models, I follow the method proposed by Gu et al. (2020) and create a latent factor model of excess returns for a given time period, denoted as $t = 1, 2, \ldots, T$. I introduce a conditional characteristic, denoted as $c^{10}_{i,t}$, which is not directly present in the latent factor model but affects its generation through a conditional dependence on a factor that is in the model. The excess returns for each stock i at time $t+1$ are modeled as:

$$r_{i,t+1} = \begin{cases} g^*_{lge}\left(c_{i_{lge},t}\right) + e_{i_{lge},t+1} & c^{10}_{i,t} \geq 0.2 \\ g^*_{sml}\left(c_{i_{sml},t}\right) + e_{i_{sml},t+1} & c^{10}_{i,t} < 0.2 \end{cases} \tag{4.11}$$

$$e_{i,t+1} = \begin{cases} \beta_{i_{lge},t} f_{t+1} + \epsilon_{i_{lge},t+1} & c^{10}_{i,t} \geq 0.2 \\ \beta_{i_{sml},t} f_{t+1} + \frac{\epsilon^S_{i_{sml},t+1}}{\epsilon_{i_{sml},t+1}} & c^{10}_{i,t} < 0.2 \end{cases} \tag{4.12}$$

$$\beta_{i,t} = \left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}, c^4_{i,t}, c^5_{i,t}, c^6_{i,t}\right), \tag{4.13}$$

where c_t is an $N \times K_c$ matrix of characteristics, f_{t+1} is an $f \times 1$ vector of factor innovations, and ϵ_{t+1} is an $N \times 1$ vector of idiosyncratic errors. I choose $f_{t+1} \sim N\left(0, 0.05^2 \times \mathbb{I}^3\right)$ and $\epsilon_{i,t+1} \sim t_5\left(0, 0.05^2\right)$. As previously described, I introduce a conditional relation between a characteristic, denoted as $c^{10}_{i,t}$, and the values of ϵ_{t+1}. In particular, for the bottom 20% of stocks as measured by the characteristic, I scale ϵ_{t+1} by ϵ^S_{t+1}. I construct a panel of cross-sectional characteristics at each time step using the following model:

$$c_{i,t}^{j} = \frac{2}{N+1} Rank\left(\bar{c}_{i,t}^{j}\right) - 1 \tag{4.14}$$

$$\bar{c}_{i,t}^{j} = \rho_j \bar{c}_{i,t-1}^{j} + \epsilon_{i,t}, \tag{4.15}$$

where $\rho \in \mathcal{U}[0.9, 1.0]$, $\epsilon_{i,t}^{j} \in N\left(0, 1-\rho_j^2\right)$, and $Rank(\cdot)$ cross-sectionally rank normalizes the simulated characteristics to be between $[-1, 1]$. This ranking is identical to the one performed on the actual stock characteristics dataset from OSAP. I explore several variations of $g^*(\cdot)$ functions with the intent to simulate various conditional relations. Further, I describe a OHE scenario. In this scenario, for the chosen characteristic that introduces the size-dependence, I first transform this into two Boolean columns. Next, I introduce two additional datasets that are effectively for the multiplication of the original dataset by the two Boolean columns. Across all scenarios, $N = 100$, $T = 120$, and $K_c = 100$ are fixed. The tenth characteristic is used to represent a non-informational characteristic in which some other model dependence is introduced. One hundred Monte Carlo samples are used. In each sample, the 100×120 row dataset is divided into three equal 100×40 datasets representing the training, validation, and test sets. Table 4.10 in Appendix 4.1 details the exact parameter configurations used in each test.

4.4.3 Simulation results

Table 4.6 presents the simulation results when an imposed conditional dependence between a selected characteristic and the prediction target is introduced into the factor DGP. The selected characteristic is used to control the degree of return predictability in two different groups. Scenarios are run with different predictability levels in the bottom 20% of stocks, varying volatility levels, no predictability, and non-linear characteristic interactions. In each scenario, three machine learning models are trained: one on the full panel, one on the top 80% based on the sorted selected characteristic, and one on the bottom 20%, simulating the effect of training on group–specific models sorted by market capitalization in the earlier empirical results.

Training independent group–specific models lead to superior results when a characteristic has higher predictability or only predicts returns for one group in the panel. For instance, when a characteristic only has statistical power for the bottom 20%, we see a significant improvement in in-sample and OOS performance when training the models separately (the IC increases from 1.44% to 3.03% for the OOS combined panel). Similarly, when the predictive ability of characteristics

for the top 80% of stocks is set to zero, the model trained only on the bottom 20% achieves a higher OOS IC than the model trained on the full cross-section. The model trained on the full cross-section achieves an OOS IC of 4.00%, whereas the model trained only on the bottom 20% reaches an IC of 8.73%. These results are perhaps not surprising. If 80% of the training data has no predictability, then these data effectively act as noise in the training process and will result in a poorer in-sample model. These results show that if there exist group–specific characteristics within the panel of training data, this can lead to inherent model biases when not accounting for these group–specific features.

To summarize, Table 4.6 shows how machine learning models can be overfit to specific groups of stocks when trained on a full cross-section with heterogeneous characteristics. Models trained on different sub-groups may perform better if different groups of stocks have different return drivers. The simulation study does not provide a conclusive resolution for the outperformance of group–specific models but confirms that the earlier empirical result on the CRSP data is not a statistical fluke but a general characteristic of training machine learning models for predicting returns. Although the improved economic performance we observed for group–specific machine learning models is significant, it comes at the cost of training three separate machine learning models. In the case of a neural network, where the common practice is to train 10 neural networks with different random seeds, the number of machine learning models required increases from 10 to 30. This requirement is not ideal, and we would preferably achieve similar performance improvements as group–specific machine learning models without the significant increase in computational cost.

Table 4.6: Machine learning model fits under simulated data generating processes

This table presents the average ICs for a series of machine learning models trained using two different approaches. The first approach (1) trains the model using all data, whereas the second approach (2) trains two separate models using a non-predictive simulated characteristic to split the sample into two using the 20th percentile cutoff and training a model on each sample separately. Each row corresponds to a different DGP, and for each DGP, 100 Monte Carlo simulations are run. An ensemble over the three separately trained models is created for each machine learning model. The table also reports the t-statistic for the null hypothesis: the mean of the distribution of ICs in (1) is less than that of ICs in (2). Bold text indicates significance at the 5% level.

	In-sample									Out-of-sample								
	All			Top 80%			Bottom 20%			All			Top 80%			Bottom 20%		
IC (%)	(1)	(2)	t-stat	(1)	(2)	t-stat	(1)	(2)	t-stat	(1)	(2)	t-stat	(1)	(2)	t-stat	(1)	(2)	t-stat
Base	7.41	8.22	**(-2.01)**	7.33	8.05	(-1.61)	7.78	10.86	**(-4.32)**	0.96	0.59	(1.28)	0.82	0.64	(0.55)	1.53	0.16	**(2.40)**
Increase in bottom 20% IC	9.66	10.97	**(-2.43)**	8.21	7.92	(0.56)	14.32	21.56	**(-6.51)**	1.88	2.02	(-0.41)	1.28	0.58	**(2.49)**	3.67	5.51	**(-2.36)**
Increase in bottom 20% volatility	7.58	9.70	**(-4.46)**	7.17	8.98	**(-3.60)**	9.10	13.37	**(-5.00)**	0.55	0.52	(0.11)	0.44	0.90	(-1.61)	0.97	-0.35	**(2.52)**
Increase in bottom 20% IC and volatility	9.20	12.05	**(-4.82)**	7.71	8.21	(-0.95)	13.81	23.20	**(-8.13)**	1.54	1.78	(-0.78)	1.10	0.58	(1.79)	2.76	4.30	**(-2.23)**
Increase in bottom 20% IC and volatility OHE	15.41	15.04	(0.46)	10.31	11.00	(-1.02)	29.05	26.95	(1.360)	2.73	2.39	(1.12)	1.07	0.73	(1.16)	6.45	5.94	(0.74)
One predictor only applies for bottom 20%	10.25	15.35	**(-7.08)**	6.58	8.13	**(-2.96)**	18.95	32.01	**(-8.90)**	1.44	3.03	**(-4.40)**	0.27	0.51	(-0.79)	3.82	7.90	**(-4.69)**
Top 80% has no predictability (1)	10.66	14.86	**(-6.51)**	6.68	7.33	(-1.53)	19.87	32.03	**(-8.91)**	1.15	3.35	**(-6.08)**	-0.35	0.38	**(-2.52)**	4.00	8.73	**(-5.61)**
Top 80% has no predictability (2)	11.15	15.78	**(-7.37)**	6.95	6.97	(-0.06)	20.84	34.93	**(-10.62)**	1.53	3.48	**(-5.32)**	0.22	-0.10	(1.22)	4.18	9.51	**(-6.23)**
Bottom 20% has exaggerated IC on same factors	57.93	86.44	**(-33.62)**	14.20	7.62	**(11.81)**	83.68	95.70	**(-20.88)**	16.96	50.35	**(-41.81)**	8.85	0.53	**(20.46)**	37.59	64.21	**(-26.5)**
Bottom 20% has exaggerated IC on different factors	90.64	90.34	(0.58)	14.48	10.19	**(7.31)**	96.26	97.40	**(-8.88)**	49.14	58.76	**(-9.51)**	0.36	0.69	(-1.24)	64.55	71.31	**(-8.72)**
Exaggerated IC on difference factors	87.00	94.35	**(-53.40)**	92.13	95.73	**(-27.10)**	62.34	95.48	**(-100.99)**	72.47	85.35	**(-46.96)**	83.21	90.66	**(-25.01)**	26.53	64.31	**(-55.13)**
Non-linear and factor interaction	6.57	7.91	**(-3.09)**	6.63	7.35	(-1.52)	6.27	11.49	**(-7.04)**	0.26	0.20	(0.25)	0.14	0.22	(-0.32)	0.77	0.16	(1.26)
Increase in bottom 20% IC via non-linear factor	7.98	19.33	**(-14.57)**	6.58	7.86	**(-2.74)**	10.84	15.32	**(-5.62)**	0.74	10.21	**(-14.81)**	0.48	0.45	(0.11)	2.10	2.09	(0.01)
Increase in bottom 20% IC via interacting factors	8.10	9.87	**(-3.82)**	6.77	7.48	(-1.46)	12.21	18.09	**(-6.34)**	1.20	1.26	(-0.23)	0.75	0.37	(1.30)	2.51	3.61	(-1.94)
Different predictors for top 80%/bottom 20%	13.26	18.43	**(-7.48)**	11.64	13.25	**(-2.24)**	18.69	33.33	**(-11.51)**	4.21	5.76	**(-4.02)**	4.49	4.40	(0.20)	4.34	9.04	**(-6.29)**
Differing predictive direction (1)	6.93	8.40	**(-3.38)**	6.94	7.70	**(-2.08)**	6.86	11.58	**(-6.18)**	0.64	0.35	(1.00)	0.81	0.43	(1.14)	-0.08	0.20	(-0.57)
Differing predictive direction (2)	7.40	9.06	**(-4.09)**	7.20	6.92	(-1.62)	8,16	14.87	**(-8.31)**	0.40	0.92	(-1.85)	0.45	0.64	(-0.65)	0.22	1.76	**(-2.31)**

4.5 Choices: features, architecture, and target

Ultimately, we are interested in the practical usage of machine learning models for asset pricing and portfolio management purposes. The behavior of machine learning models using simulated factor DGPs provides insights into the underlying mechanics but is limited in practical relevance. Through the simulation exercise, I found that neural network models can overfit groups of assets within the training dataset. Using this insight, I now conduct empirical experiments to investigate how machine learning design decisions affect model performance and which design choices can reduce this group–specific overfitting. Specifically, I focus on three critical areas of model design decisions: features, architecture, and target. I make stylized choices within each category and analyze their impact on stock-level return predictions and portfolio performance. I do not aim to cover every possible modeling decision but rather to explore the common representative choices observed in literature and additional cases related to the group–specific model results. I exclusively focus on the NN3 model, given the higher propensity for overfitting of neural network architectures have for overfitting compared with tree-based models.

4.5.1 Features

The input features are the data the model uses to learn the functional form $g(\cdot)$ of $\mathbb{E}(r_{i,t+1})$ that maximizes predictive power. Numerous choices need to be made during input feature design, which is often known as "feature engineering" in the machine learning lexicon. I analyze two critical decisions, feature selection and feature normalization, and examine how these choices affect the quality of excess return predictions and long–short portfolios. The choice of which features to provide a machine learning model is often underappreciated, particularly in asset pricing. Given the high complexity associated with typical neural network models, the inclusion (or exclusion) of specific features can impact the resulting predictions in unexpected ways.

I first employ ex-post feature importance selection. I train machine learning models following the earlier approach. I then calculate the top 10 most important features across the full OOS period. Using these top features, I create two scenarios: "Drop top," where I drop the top 10 features from the input feature set and retrain each model, and "Only top," where I only use the top 10 features to train each model. Note that this approach introduces a degree of look-ahead bias, as I use feature importance from OOS periods during earlier periods of machine learning training. This decision is akin to the standard approach to asset pricing, where a set of features or asset pricing anomalies that were not yet discovered are used to train

and evaluate models. The list of top characteristics for each model can be found in Appendix 4.6.

The second choice involves creating a true OOS input feature set, where I use the reported year of characteristic discovery from OSAP. If a characteristic was published in the literature in 1997, this characteristic is only used for training after January 1998. The available set of characteristics annually increases as the training window expands each year.

The final design choices explore the interaction between the continuous Size characteristic and the other input features. In the earlier results, where we saw the outperformance of group–specific models, the training dataset is split based on the Size characteristic. However, it is natural to ask why the machine learning model did not independently capture this effect. I create two new datasets to test alternative approaches for capturing the interaction between Size and model predictive power. In the first scenario, I use one-hot encoding (OHE) to create three additional dummy columns for Size, indicating whether each stock is in the Top 30%, Middle 40%, or Bottom 30% of Size. In the second scenario, I take the Size dummies from the previous scenario and compute the inner product of each dummy variable with the existing input variable set, thereby creating three additional datasets of equal size as the base dataset but with non-zero values only for rows corresponding to the given Size group. These three additional datasets are concatenated with the original dataset and then used as the input feature set.

Table 4.7 shows the results of the feature choices applied to train the NN3 model using all data and size-groups. The table reports the average IC, average annualized top-minus-bottom return, average annualized portfolio volatility, average Sharpe ratio (SR), largest one-month portfolio loss (DD1M), maximum portfolio drawdown (MaxDD), average monthly one-way portfolio turnover (TO), FF6 alpha and corresponding t-statistic, FF6 and HXZ break-even transaction costs, and three test statistics: DM, IC, and Ledoit-Wolfe Sharpe ratio. The last three columns compare each model against the baseline feature set trained on all CRSP data.

None of the feature permutation models consistently outperform the base input feature set trained in size–specific groups. Although some models have higher break-even transaction costs, they are often associated with lower ICs and average returns. For instance, the model trained on all CRSP stocks using the top 10 features has the highest FF6 break-even transaction cost and DM test statistic. However, its IC is significantly lower than that of other models. This model also has more embedded look-ahead bias, making it uncertain whether the model will continue to outperform.

Overall, none of the feature design choices resolve the empirical observation, as we still find that group–specific models outperform.

4.5.2 Architecture

A neural network's architecture has effectively unlimited design possibilities, and there is no fixed formula for optimal architecture design. Instead, the approach is typically based on trial and error across various design choices. Guidance can be drawn from machine learning research in other fields; however, applying it to return prediction is not always straightforward. As Israel, Kelly and Moskowitz (2020) show, using machine learning in finance is challenging due to the small data problem and low signal-to-noise ratios. I examine several commonly made assumptions and decisions when designing neural networks, such as batch normalization and dropout. First, I investigate different loss functions used in the Adam optimizer, including the Huber loss function and the mean absolute error (MAE) loss function. The choice of loss function is critical, as it is what the model is ultimately trying to optimize. The Huber loss function is specified as:

$$L_H = \begin{cases} \frac{1}{2}\left(r_{i,t+1} - \hat{r}_{i,t+1}\right)^2 & for\, |r_{i,t+1} - \hat{r}_{i,t+1}| \leq q_{0.999} \\ \left(q_{0.999} \times |r_{i,t+1} - \hat{r}_{i,t+1}| - \frac{1}{2} q_{0.999}\right) & otherwise, \end{cases} \tag{4.16}$$

where the δ parameter is set to the 99.9th quantile of excess returns and is designed to regularize the neural network algorithm, reducing the impact of extreme small stock outliers on the overall sample.

Table 4.7: Neural network portfolio performance with different input feature choices

This table compares the OOS VW portfolio performance of the NN3 model with different input feature choices. The table reports portfolio performance and three test statistics: DM, IC, and Ledoit Wolfe Sharpe ratio. The universe column indicates whether the model was trained once on all stocks (All) or separately on large, mid, and small stocks (Size). The feature column specifies the approach used to determine the input feature set used to train the NN3 model, including Base, Drop top, Only top, Survivorship bias, OHE size features, and OHE size interacted. In Base, I use cross-sectional standardization over $[-1, +1]$ to create the input feature set. In Drop top, I drop the top 10 features as determined by the average feature importance across the Base model and retrain the models from scratch. In Only top, I keep only the top 10 important features from the Base model and retrain the models. Survivorship bias only uses features that were known at each annual training date. OHE size features include three binary columns for large/mid/small capitalization companies. In OHE size interacted, I take the inner product between the Base feature set and three size dummies to produce three additional input feature sets that contain non-zero values only for each size category.

Universe	Feature	IC (%)	Ret. (ann.)	Vola. (ann.)	SR (ann.)	DD1M (%)	MaxDD (%)	TO (%)	FF6 α	FF6 $t(\alpha)$	FF6 break-even t-cost	HXZ break-even t-cost	DM t-stat	IC t-stat	LW t-stat
Size	Survivorship bias	3.08	11.0	12.5	0.88	24.1	51.5	125	8.7	2.8	29.00	36.99	-1.89	-3.04	0.38
Size	Base	5.53	15.0	17.4	0.87	15.6	44.8	134	12.3	4.16	38.12	35.60	1.70	2.11	0.73
All	Drop top	3.72	10.8	13.2	0.81	16.3	45	135	9.5	4.11	29.36	34.55	0.46	-1.29	0.30
All	Base	4.63	11.6	15.5	0.75	19.2	46.9	137	11.2	4.05	34.21	31.23	–	–	–
Size	Drop top	3.95	12.6	16.9	0.75	17.4	62.4	124	12	3.41	40.23	35.91	1.30	-1.69	0.07
Size	Only top	4.66	12.0	16.7	0.72	23.0	62.3	152	9.8	2.9	26.87	29.87	1.84	0.44	-0.21
All	OHE size features	4.24	10.6	15.5	0.69	22.2	39.9	141	9.8	3.42	28.95	36.97	-0.14	0.41	-0.53
All	OHE size interacted	4.05	9.9	14.9	0.66	16.1	63.3	144	9.3	3.33	26.83	31.29	-2.78	-2.04	-0.39
All	Only top	3.39	12.8	20.1	0.64	17.7	30.1	151	9.3	3.24	25.59	25.66	2.42	-1.58	-0.64
All	Survivorship bias	2.18	6.6	12.7	0.52	16.0	50.5	129	6.0	1.88	19.28	19.37	-2.45	-2.84	-1.12

Neural networks are canonically known to create non-linearity and interaction between features without having to explicitly model them. Neural network models achieve this using activation functions that transform the outputs of the hidden layers. I use the leaky rectified linear activation function (ReLU) and exponential linear unit (ELU) instead of the standard ReLU between hidden layers. The ReLU activation function can experience the dying ReLU problem (Lu, Shin, Su and Karniakdakis, 2020), where the output of the ReLU layer is constant for all inputs. Leaky ReLU and ELU activation functions, which have been suggested as alternative options, avoid this issue. I also consider changes to the regularization layers used. In the base model, only L1 regularization is used in the hidden layers, and the L1 penalty is tuned as a hyperparameter. I investigate using only an L2 regularization layer, both L1 and L2 regularization layers, and no regularization layers. Each option requires tuning additional hyperparameters to determine the appropriate L1 and L2 penalties.

Instead of using a three-layer neural network architecture, I test alternative structures: a wider neural network with four hidden layers, having (1054, 512, 258, 128) units, a deeper neural network with eight hidden layers, having (256, 128, 64, 32, 16, 8, 4, 2) units, and an expanding neural network with six hidden layers, having (32, 64, 128, 256, 16, 8) units.

I finally test removing batch normalization, removing early stopping, and introducing dropout to the hidden layers at a rate of 0.2. These architectural decisions are often used to regularize the neural network and prevent overfitting. Dropout is a widely used regularization technique for training neural networks, where connections between hidden layers are randomly removed. This process encourages the model to rely on a diverse range of node dependencies, helping to reduce overfitting.

Table 4.8: Neural network portfolio performance with different model architecture choices

This table presents the OOS VW portfolio performance for the NN3 model under different machine learning architecture choices. The universe column indicates whether the model was trained once on all stocks (All) or separately on large, mid, and small stocks (Size). The feature column specifies the approach to determine the architecture used to train the NN3 model. The loss function used in the ADAM optimizer is changed from the MSE to Huber or MAE. The activation function used in the hidden layers is changed from the ReLU to leaky ReLU or ELU. The use of L1 and L2 regularization penalties in the hidden layers is changed from the Base L1 penalty only to no penalty, only L2 penalty, and both L1 and L2 penalties. The dimensions and number of hidden layers used are also changed, including a Deeper NN with eight hidden layers, a Wider NN with four hidden layers, and an Expanding NN with six hidden layers. I also test cases where no batch normalization or early stopping is used and where dropout is used in the hidden layers at a rate of 0.2.

Universe	Feature	IC (%)	Ret. (ann.)	Vola. (ann.)	SR (ann.)	DD1M (%)	MaxDD (%)	TO (%)	FF6 α	FF6 $t(\alpha)$	FF6 break-even t-cost	HXZ break-even t-cost	DM t-stat	IC t-stat	LW t-stat
Size	No batch norm	6.08	28.6	16.2	1.76	19.3	28.4	132	26.4	7.4	83.00	83.28	3.10	3.67	4.55
Size	No early stop	4.97	21.8	17.3	1.26	17.4	30.5	130	21.2	6.1	67.92	61.24	-1.18	1.31	2.31
Size	Dropout	6.28	23.3	20.4	1.14	17.0	64.4	135	21.5	4.5	66.34	61.48	3.96	3.42	1.87
Size	Leaky ReLU	5.32	17.5	16.4	1.07	21.8	50.7	135	16.2	4.97	50.07	47.26	1.21	1.50	1.54
Size	ELU	5.64	18.0	17.3	1.04	22.9	33.6	137	14.6	5.16	44.44	48.09	1.69	2.15	1.45
Size	L1 & L2 penalty	5.13	15.9	15.2	1.04	14.1	37.4	130	15.2	5.12	48.81	44.61	2.29	1.50	1.28
Size	Huber loss	5.92	18.1	17.9	1.01	23.8	47.9	133	15.9	4.57	50.00	45.35	2.24	2.49	1.27
Size	L2 penalty	5.01	15.8	16.3	0.97	22.4	55.1	135	13.9	3.98	43.06	38.41	1.06	0.80	1.06
All	ELU	4.26	14.6	15.6	0.94	19.2	35.5	140	14.8	5.29	44.01	43.81	-0.87	-1.14	0.96
Size	Base	5.53	15.0	17.4	0.87	15.6	44.8	134	12.3	4.16	38.12	35.60	1.70	2.11	0.73
Size	Wider NN	3.33	12.5	14.7	0.86	15.0	43.8	111	14.2	4.32	53.02	47.41	-3.70	-2.33	0.58
All	No early stop	3.56	13.7	16.5	0.83	18.5	29.1	141	12.1	4.36	35.82	33.89	-3.14	-2.42	0.39
All	MAE Loss	4.99	17.2	21.0	0.82	28.2	57.2	123	16.3	4.1	55.16	50.89	-0.76	-0.8	0.25
Size	Expanding NN	4.57	12.9	15.9	0.82	16.3	51.0	154	7.7	3.76	20.89	17.84	1.95	0.98	0.23
All	No batch norm	4.59	9.4	12.1	0.78	10.9	12.3	119	8.4	4.08	29.64	23.64	1.81	-1.43	0.12
Size	No L1 & L2	4.00	9.5	12.4	0.77	12.2	26.0	177	5.5	3.00	13.00	8.80	-1.39	-1.32	0.07
Size	MAE loss	5.26	17.7	23.6	0.75	24.9	74.9	116	14.8	3.39	53.17	51.90	-0.80	-0.27	0.04
All	Base	4.63	11.6	15.5	0.75	19.2	46.9	137	11.2	4.05	34.21	31.23	–	–	–
All	L2 penalty	4.56	9.8	13.7	0.72	19.1	33.1	138	9.7	3.58	29.34	28.87	0.54	-0.54	-0.22
All	Dropout	4.01	11.2	16.1	0.70	17.1	41.5	141	10	3.15	29.66	32.91	2.61	-1.43	-0.35
All	Huber loss	4.26	9.8	14.4	0.68	19.2	42.8	140	8.7	3.18	25.77	25.48	-0.37	-1.56	-0.55
All	L1 & L2 penalty	4.44	9.7	15.2	0.64	18.4	40.9	128	9.2	3.04	30.05	29.48	-0.61	-0.9	-0.68
All	Deeper NN	3.63	10.9	17.3	0.63	21.3	62.1	129	10.9	3.16	35.00	35.98	-0.72	-2.19	-0.71
All	Leaky ReLU	3.88	9.2	15.5	0.59	20.8	58.3	140	8.3	2.7	24.78	23.59	-0.70	-2.23	-0.92
All	Expanding NN	3.53	9.0	15.3	0.59	20	49.9	161	5.5	2.51	14.19	6.20	-2.79	-3.74	-0.93
All	No L1 and L2	3.58	6.7	12.0	0.56	21.9	46.6	143	6.6	2.88	19.13	18.44	-1.66	-2.11	-0.95
Size	Deeper NN	2.96	6.7	12.3	0.54	11.5	43.0	113	5.6	2.46	20.73	20.72	0.54	-1.93	-1.07
All	Wider NN	2.95	7.3	14.5	0.50	18.0	47.3	141	5.2	1.86	15.29	15.85	-4.73	-1.65	-1.48

Table 4.8 presents the results of the changes to the neural network architecture. Notably, the model trained without batch normalization demonstrated a significant increase in Sharpe ratio from 0.86 to 1.76 compared with the base group–specific model. This improvement is consistent across various metrics, including IC, top-minus-bottom returns, break-even transaction costs, and stock-level return prediction tests. Using more complex neural network architectures does not necessarily lead to better performance, particularly for the deeper neural network architecture. This result is counter-intuitive to the commonly accepted practice that deeper neural networks are superior. However, Gu et al. (2020) also find that shallow neural networks outperform. The results of altering the neural network architecture do not provide a convincing explanation for why group–specific models outperform those trained on all samples. Instead, these results highlight the potential for substantial improvements in machine learning predictions by modifying common design choices, even those widely accepted, such as always employing batch normalization. The strength of neural networks is the inherent complexity they can flexibly model; however, this flexibility comes at the cost of many design levers that can be pulled, and the finance literature has only begun to scratch the surface of optimal neural network design in asset pricing.

4.5.3 Target

The final modeling choice I explore is the target variable design. Conventionally, in empirical asset pricing, raw excess returns are used without adjustment. I experiment with alternative approaches, such as classification, instead of regression. In this analysis, I do not consider residual returns, as doing so would alter the prediction problem by conditioning it on the factors used for residualization. For instance, removing the component of excess returns explained by the FF3 model changes the underlying prediction problem. I focus on design choices that preserve the fundamental prediction problem as much as possible.

The first set of changes involves subtracting the median excess return. In the "Median" case, the cross-sectional median excess return is subtracted from each stock's monthly excess return. In the "Median size group" case, the corresponding cross-sectional median excess return within the three size groups is subtracted from each stock's monthly excess returns. The second set of changes involves rank normalizing returns. In the "Rank" case, stocks are cross-sectionally ranked each month and then scaled between $[-1, 1]$. In the "Rank in size" case, stocks are cross-sectionally ranked within size groups each month and then scaled between $[-1, 1]$. In the "Winsorize" case, cross-sectional excess returns are winsorized each month at the 1% level.

In addition to testing a regression problem, I explore classification. In particular, the "Base classifier" scenario creates a binary classification problem: to classify whether excess returns are positive or negative. Stocks are cross-sectionally sorted using the classification probability. In the "Classifier median" case, excess returns are classified as greater than or equal to the monthly median excess return (positive) or less than the monthly median excess return (negative). Last, in the "Classifier median size" case, excess returns are classified as greater than or equal to the monthly median excess return within the corresponding size groups (positive) or less than the monthly median excess return within size groups (negative).

Table 4.9 presents the results from the target variable changes. We observe consistent improvements across most scenarios compared with the baseline cases. Notably, in the "Median size group" scenario, the group–specific and full-trained models achieve comparable results. This finding could help explain the empirical observation where group–specific models outperform models trained on the full cross-section. Removing the component of a stock's return associated with the size group from the prediction problem aligns the model's goal with predicting cross-sectional return distributions rather than predicting whether small stocks will outperform large stocks. Regularizing the target variable by removing median excess returns or rank normalizing leads to better fits of the machine learning models. These results suggest that across the three design choices we have explored, careful calibration of the target variable has the most significant impact on the subsequent portfolio performance.

Table 4.9: Neural network portfolio performance with different target choices

This table presents the OOS VW portfolio performance for the NN3 model under different prediction target choices. The universe column indicates whether the model was trained once on all stocks (All) or separately on large, mid, and small stocks (Size). The feature column specifies the approach to determine the target used to train the NN3 model. The Base case uses the raw stock-level excess returns obtained from CRSP. In each case, the excess stock returns are the starting point, and they are adjusted each month. The different target choices include the following: Rank, where excess returns are cross-sectionally ranked and scale over $[-1, +1]$; Rank in Size, where excess returns are cross-sectionally ranked within three separate size categories and then scaled over $[-1, +1]$; Median, where the corresponding monthly cross-sectional median excess return is subtracted from the excess returns; Median size group, where the corresponding monthly cross-sectional median excess return within size groups is subtracted from the excess returns; Winsorize, where excess returns are winsorized at the 1% level; Base classifier, where excess returns greater than or equal to zero are labeled as 1, and excess returns less than zero are labeled as 0; Classifier Median, where excess returns greater than or equal to the monthly cross-sectional median excess return are labeled as 1, and otherwise 0; and Classifier Median Size, where excess returns greater than or equal to the monthly cross-sectional median excess returns within size groups are labeled as 1, and otherwise 0. The DM t-statistic is undefined in cases where the regression target is altered, as the excess return predictions are not directly comparable with the Base model, which makes predictions of raw excess return.

Universe	Feature	IC (%)	Ret. (ann.)	Vola. (ann.)	SR (ann.)	DD1M (%)	MaxDD (%)	TO (%)	FF6 α	FF6 $t(\alpha)$	FF6 break-even t-cost	HXZ break-even t-cost	IC t-stat	LW t-stat
Size	Median size group	5.72	23.3	15.8	1.48	13.1	23.7	138	20.9	6.71	63.20	61.96	2.52	3.37
All	Median size group	6.05	24.1	18.7	1.29	20.7	35.1	137	22.3	6.29	67.78	64.50	3.04	2.56
Size	Classifier median size	4.97	18.9	16.2	1.17	17.5	42.0	146	17.8	5.64	50.71	46.59	0.37	1.92
Size	Median	5.85	19.3	16.6	1.17	14.0	38.2	136	17.3	5.34	52.90	53.23	2.73	2.06
Size	Rank in size	6.29	19.3	16.6	1.16	24.2	39.6	144	18.6	6.59	53.85	47.36	1.31	1.95
All	Classifier median size	5.60	21.2	19.0	1.12	19.5	40.0	145	19.4	5.39	55.71	53.02	-0.47	1.69
All	Rank	6.32	24.5	22.0	1.11	30.4	49.3	128	23.5	5.39	76.56	72.58	1.11	1.87
Size	Rank	6.31	18.9	17.3	1.09	18.4	49.3	141	17.3	4.80	50.84	48.77	1.38	1.73
Size	Classifier median	5.65	16.1	15.0	1.07	18.6	33.6	142	15.2	4.89	44.39	42.68	-0.37	1.52
All	Median	6.46	23.0	22.2	1.04	29.6	64.6	141	20.6	4.49	61.11	55.47	2.59	1.46
All	Rank in size	6.43	20.7	20.7	1.00	22.2	43.9	139	19.3	4.96	57.97	57.64	1.31	1.4
Size	Winsorize	5.65	17.3	17.7	0.98	13.3	37.7	136	14.6	4.56	44.71	46.57	1.77	1.19
All	Base classifier	3.97	18.4	19.1	0.96	29.9	65.2	130	17.6	3.87	56.49	52.33	-2.78	0.95
Size	Base	5.53	15.0	17.4	0.87	15.6	44.8	134	12.3	4.16	38.12	35.60	2.11	0.73
All	Winsorize	4.67	11.1	13.6	0.82	10.7	29.9	139	9.9	3.97	29.75	27.11	-0.31	0.27
All	Base	4.63	11.6	15.5	0.75	19.2	46.9	137	11.2	4.05	34.21	31.23	–	–
Size	Base classifier	3.45	13.9	19.1	0.73	21.4	64.9	124	12.9	3.22	43.35	39.53	-2.64	-0.15

4.5.4 A partial resolution

Figure 4.6 depicts the top-minus-bottom returns and ICs of training on all stocks versus training on size–specific groups with an adjusted target of returns above the median return within size groups. I present this as a partial resolution to the empirical anomaly I found of superior machine learning model performance when training on smaller group–specific datasets. Implicit regularization obtained through training on size–specific groups is achieved by explicitly regularizing the target variable without incurring the computational cost of training three independent models. It is non-trivial to delineate whether this improvement comes from imparting the economic prior on group–specific models or simply removing the market return from the target variable. In either case, I conclude that performing some form of regularization on the target variable is prudent to achieve superior predictive performance when training machine learning models for return prediction.

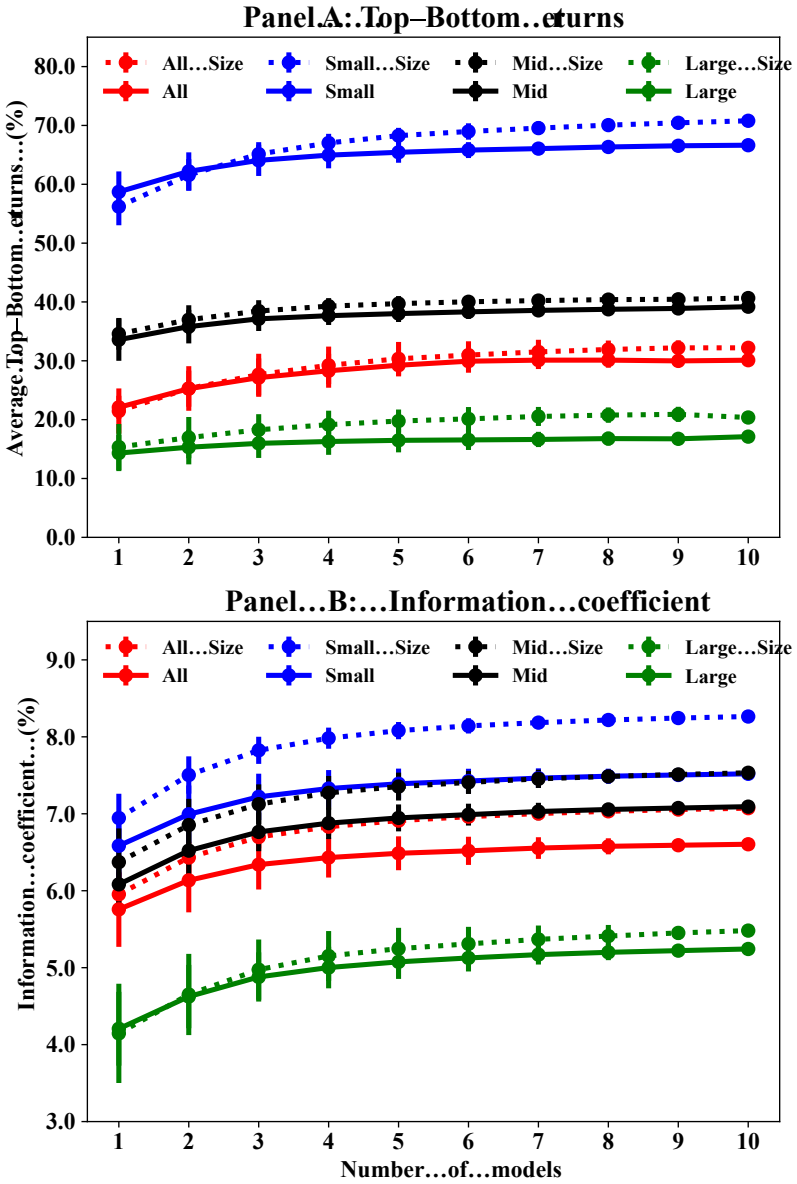

Figure 4.6: Effect of model averaging across neural networks with a regularized target variable

This figure presents the average annualized top-minus-bottom portfolio return (Panel A) and average IC (Panel B) for the NN3 model when ensembling identically trained models with different random seeds. The figure presents the results when the cross-sectional monthly excess return within size-groups is subtracted from the corresponding stocks' monthly excess returns. The x-axis represents the number of models included in the ensemble. The solid lines present the results when training the NN3 model once using all stocks. The dashed lines present the results when training the NN3 model in three separate group–specific models and concatenating the results. Ten models are trained using an identical approach but with different randomized seeds for each size group. A quasi-Monte Carlo approach is used to measure the effect of ensembling on the uncertainty of performance statistics. This approach is detailed in Section 4.3.3. This figure presents the average and standard errors of top-minus-bottom returns and ICs of the portfolios formed from these ensemble predictions.

4.6 Conclusion

Finance literature has only just begun to explore the application of machine learning models for predicting cross-sectional stock returns. There is no standard modeling framework for comparing results across different studies. The high dimensionality of choices associated with machine learning modeling in asset pricing results in a high level of complexity in attributing performance gains related to changes to machine learning modeling approaches. This study contributes to the field by training group–specific machine learning models and demonstrating superior predictive and portfolio performance compared with a model trained on the full dataset. By investigating various machine learning design choices, I reveal that a lack of regularization of the target variable primarily drives the outperformance of group–specific machine learning models. By implementing target variable regularization, the performance gains associated with group–specific machine learning models can be achieved at lower computational complexity.

I assess the impact of various design choices on prediction outcomes, including feature selection, model architecture, and target variable regularization. The findings indicate that regularizing the target variable—total excess returns—significantly contributes to the observed outperformance of the group–specific models. My results emphasize that the lack of consistency in the literature regarding model design choices hinders the advancement of machine learning in finance and obscures comparative analysis across studies. In this case, should the improvement in model performance be attributed to the economic rationale behind the group–specific modeling decision or to the regularization of the target variable?

By shedding light on the impact of these design choices on stock portfolio formation, machine learning models can more effectively be employed in financial settings. I encourage future research to explore and standardize design choices to promote more robust and reliable machine learning applications in finance. A standardized approach will facilitate the development of a more unified body of knowledge and enable more precise comparisons of results across studies, ultimately enhancing the potential for machine learning to support financial decision-making and portfolio management.

Appendix 4.1. Data generating process for simulations

Table 4.10: Simulation parameters

This table presents the simulation parameters used for each scenario presented in Table 4.6, and the method is described in Appendix 4.6.

Scenario	$g^*_{lge}(\cdot)$	$g^*_{sml}(\cdot)$	ϵ_{t+1}	ϵ^S_{t+1}	θ_{lge}	θ_{sml}	OHE
Base	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(0.02, 0.02, 0.02)$	N
Increase in bottom 20% IC	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(0.04, 0.04, 0.04)$	N
Increase in bottom 20% volatility	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.02, 0.02, 0.02)$	$(0.02, 0.02, 0.02)$	N
Increase in bottom 20% IC and volatility	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.02, 0.02, 0.02)$	$(0.04, 0.04, 0.04)$	N
Increase in bottom 20% IC and volatility OHE	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.02, 0.02, 0.02)$	$(0.04, 0.04, 0.04)$	Y
One predictor only applies for bottom 20%	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.02, 0.02, 0)$	$(0.02, 0.02, 0.30)$	N
Top 80% has no predictability (1)	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0, 0, 0)$	$(0.02, 0.02, 0.30)$	N
Top 80% has no predictability (2)	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.01, 0.01, 0)$	$(0.02, 0.10, 0.30)$	N
Bottom 20% has exaggerated IC on same factors	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(2, 2, 2, 2)$	N
Bottom 20% has exaggerated IC on different factors	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}, c^4_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}, c^4_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0, 0, 0, 0.02)$	$(2, 2, 2, 0)$	N
Exaggerated IC on difference factors	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}, c^4_{i,t}, c^5_{i,t}, c^6_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}, c^4_{i,t} c^4_{i,t}, c^5_{i,t}, c^6_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0, 0, 0, 2, 2, 2)$	$(2, 2, 2, 0, 0, 0)$	N
Non-linear and factor interaction	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(0.02, 0.02, 0.02)$	N
Increase in bottom 20% IC via non-linear factor	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(0.02, 0.30, 0.02)$	N
Increase in bottom 20% IC via interacting factors	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(\left(c^1_{i,t}\right)^2, c^1_{i,t} \times c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(0.02, 0.02, 0.30)$	N
Different predictors for top 80%/bottom 20%	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.075	$(0.02, 0.10, 0)$	$(0.02, 0, 0.30)$	N
Differing predictive direction (1)	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(-c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(-0.02, 0.02, 0.02)$	N
Differing predictive direction (2)	$\left(c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{lge}$	$\left(-c^1_{i,t}, c^2_{i,t}, c^3_{i,t}\right)\theta_{sml}$	0.05	0.05	$(0.02, 0.02, 0.02)$	$(-0.10, 0.02, 0.02)$	N

Appendix 4.2. Machine learning models and hyperparameters

For the baseline neural network models, a batch size of 10,000, 100 epochs, early stopping with early stopping patience of 5 and an early stopping split of 0.25 are used. I use the ReLU as the hidden layer activation function, with a linear activation function in the output layer. I also use the Adam optimizer, where the learning rate η is a hyperparameter, and an MSE loss function is used. Further, I use between 1–5 hidden layers, with the architecture for the five-hidden-layers network being (32,16,8,4,2) units and that for the one hidden layer network being (32) units. I also use batch normalization layers between two hidden layers.

Table 4.11: Machine learning model hyperparameters
This table shows the values used in the hyperparameter tuning of all models. The table lists the hyperparameters used for regularized linear models, tree-based networks, and neural networks. For the regularized linear models, α is constant and multiplies the penalty terms. For the tree-based networks, depth controls the maximum allowable depth, NTrees is the total number of trees used in the estimator, and Colsample is the number of columns taken when splitting. For GBRT, Subsample is the percentage of rows used in each weak learner, and η is the learning rate. For the neural network, L1 is a scalar on the penalty in the loss function, and η is the learning rate used in the optimizer.

OLS	ENET	RF	GBRT	NN
Huber loss	$\alpha \in [0.001, 0.01, 0.1, 1.0]$ L1 ratio = 0.5	$Depth \in [3, 6]$ $NTrees = 300$ $Colsample \in [0.01, 0.1, 0.2, 1.0]$	$Depth \in [3, 6]$ $NTrees = 250$ $Subsample \in [0.3, 0.5, 1.0]$ $\eta \in [0.01, 0.1]$ $Colsample \in [0.3, 0.5, 1.0]$	$L1 \in [0.0001, 0.001, 0.01]$ $\eta \in [0.001, 0.01]$

Appendix 4.3. Stock characteristics

Table 4.12: Top ten important features across the four machine learning models
This table contains the top ten most important features for each NN3 model that was trained on different data samples.

Rank	All	Large	Mid	Small
1	STreversal	STreversal	STreversal	STreversal
2	Size	TrendFactor	High52	Size
3	High52	High52	IndRetBig	High52
4	MomRev	BM	IdioVol3F	IndRetBig
5	AccrualsBM	IdioVol3F	MaxRet	Mom6m
6	IndRetBig	IdioRisk	AnnouncementReturn	BM
7	RDcap	AM	BM	MomSeason
8	CHForecastAccrual	CF	IdioRisk	IdioVolAHT
9	BM	AnnouncementReturn	IndMom	AM
10	RDS	IndRetBig	TrendFactor	MaxRet

Appendix 4.4. Additional results

This section presents several additional results for the group–specific machine learning portfolios. Table 4.13 presents the results for EW top-minus-bottom portfolios. The results are like the main VW results that were presented, but the magnitude of improvements is smaller. This result is expected, as we observed that the machine learning training process results in a bias toward minimizing prediction errors when predicting the returns of small stocks. When portfolios are EW, this implicitly increases the relative importance of small stocks in the top and bottom portfolios. Thus, there is less room to improve the overall portfolio performance, as the baseline machine learning model is already biased toward smaller stocks and reducing this bias does not significantly improve the estimated portfolio performance. Table 4.14, Table 4.15, and Table 4.16 emphasize this result by focusing on the portfolio performance within the three size–specific universes. The largest performance improvements are clearly found for the top 30% of stocks, where the average annualized top-minus-bottom portfolio return increases from 11.9% to 19.7% and the Sharpe ratio increases from 0.88 to 1.25. For the bottom 30% of stocks, the top-minus-bottom portfolio return increases from 55.7% to 66.8%, but the Sharpe ratio decreases.

Table 4.13: Out-of-sample performance of equal-weighted machine learning portfolios under different training regimes

This table presents EW top-minus-bottom portfolio statistics for machine learning models trained using three approaches: Full, Size, and Ensemble. The Full model follows the standard approach using all available stocks, whereas the Size model trains three separate models for large, mid, and small stocks and then concatenates the predictions to form portfolios. The Ensemble model is the average of the return predictions from the Full model and Size model. The table presents the performance statistics of these portfolios, including the annualized mean, standard deviation, Sharpe ratios, maximum drawdown, maximum one-month loss, average monthly one-way turnover, annualized FF6 alpha and t-statistic, and break-even transaction costs under the FF6 and HXZ risk models. The OOS period is from January 1987 to December 2021.

Metric	Model	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Portfolio return (ann. %)	Full	24.0	21.2	21.5	27.2	42.5	41.6	41.6	42.8	42.1	45.7	44.5
	Size	24.9	39.4	12.8	32.1	50.4	50.9	49.5	49.6	46.0	50.8	51.7
	Ensemble	25.9	40.0	19.0	36.4	51.1	50.9	50.0	49.9	48.6	52.1	51.8
Volatility (ann. %)	Full	18.1	11.3	13.3	19.4	14.2	13.8	14.2	13.7	14.6	15.1	14.4
	Size	20.0	14.8	14.8	16.2	13.8	14.3	14.4	14.4	14.4	15.8	14.6
	Ensemble	20.6	14.2	15.8	18.9	15.0	14.9	15.0	14.8	15.2	16.0	15.2
Sharpe ratio	Full	1.32	1.87	1.61	1.4	2.99	3.02	2.92	3.12	2.88	3.03	3.09
	Size	1.24	2.65	0.86	1.98	3.64	3.56	3.43	3.44	3.21	3.21	3.53
	Ensemble	1.26	2.83	1.2	1.92	3.42	3.41	3.33	3.38	3.2	3.27	3.41
Max. drawdown (%)	Full	87.5	19.2	18.1	48.6	9.8	10.3	10.8	10.0	13.0	23.3	8.8
	Size	86.3	8.5	42.0	12.7	14.5	12.6	13.3	11.7	13.5	14.0	12.1
	Ensemble	90.9	10.7	40.0	11.1	13.3	10.7	10.4	10.2	13.6	18.2	13.5
Max. 1M loss (%)	Full	26.4	15.8	11.4	15.5	7.4	10.3	10.3	7.4	13.0	16.4	8.4
	Size	31.9	7.1	9.2	12.7	6.5	9.2	6.7	8.0	6.3	8.3	6.2
	Ensemble	28.7	10.7	9.0	6.4	7.6	10.7	8.2	6.0	7.5	9.9	8.0
Monthly one-way turnover (ann. %)	Full	112.1	113.0	99.5	138.1	129.2	129.2	129.4	128.1	126.3	134.1	129.9
	Size	115.9	138.8	104.9	131.9	126.6	125.9	125.6	125.3	121.9	128.2	125.7
	Ensemble	115.2	146.7	99.9	135.5	130.2	129.3	128.8	127.8	125.6	133.6	130.4
FF6 α (ann. %)	Full	22.6	20.1	20.8	23.8	40.3	39.6	39.7	40.4	39.3	42.7	42.1
	Size	24.1	37.9	10.7	30.5	48.5	48.5	46.9	47.2	43.8	48.7	49.4
	Ensemble	24.8	38.8	16.4	33.1	48.6	48.5	47.5	47.2	45.8	49.4	49.3
FF6 t-stat	Full	5.46	7.43	7.94	5.56	12.2	12.35	11.9	12.35	11.32	12.15	12.2
	Size	5.08	9.99	4.22	9.29	12.95	13.43	13.07	12.7	12.08	12.09	12.91
	Ensemble	5.1	11.1	5.69	8.43	12.8	13.51	12.8	12.68	12.12	12.63	12.81
FF6 break-even cost (bps)	Full	84.0	74.3	87.1	71.7	130.0	127.7	127.8	131.5	129.7	132.6	134.9
	Size	86.8	113.7	42.6	96.4	159.7	160.7	155.8	156.8	149.7	158.1	163.6
	Ensemble	89.7	110.2	68.3	101.9	155.5	156.2	153.6	153.9	151.8	154.0	157.5
HXZ break-even cost (bps)	Full	81.0	72.0	87.5	70.3	124.4	120.7	122.3	127.8	126.3	127.2	129.9
	Size	84.8	108.0	40.6	87.7	153.2	154.6	150.0	151.5	142.9	149.7	157.0
	Ensemble	86.8	104.8	67.1	96.2	149.7	149.6	147.5	149.0	146.2	146.8	151.6

Table 4.14: Out-of-sample performance of machine learning portfolios under different training regimes in the top 30% of stocks

This table presents VW top-minus-bottom portfolio statistics for machine learning models trained using three approaches: Full, Size, and Ensemble. Results are presented within the large stock universe, i.e., the top 30% of stocks. The Full model follows the standard approach using all available stocks, whereas the Size model trains three separate models for large, mid, and small stocks and then concatenates the predictions to form portfolios. The Ensemble model is the average of the return predictions from the Full model and Size model. The table presents the performance statistics of these portfolios, including the annualized mean, standard deviation, Sharpe ratios, maximum drawdown, maximum one-month loss, average monthly one-way turnover, annualized FF6 alpha and t-statistic, and break-even transaction costs under the FF6 and HXZ risk models. The OOS period is from January 1987 to December 2021.

Metric	Model	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Portfolio return (ann. %)	Full	9.4	5.3	8.5	3.2	10.6	11.9	10.5	11.2	11.0	11.6	11.9
	Size	11.7	14.7	8.9	11.5	15.3	21.5	18.8	18.5	13.2	20.1	19.7
	Ensemble	11.7	12.7	11.1	12.1	14.8	17.9	17.4	15.8	16.0	17.2	17.5
Volatility (ann. %)	Full	13.0	12.2	12.5	15.5	12.4	13.1	13.5	14.2	14.9	13.1	13.6
	Size	16.7	11.9	13.3	15.1	14.3	15.2	15.8	14.8	13.6	15.6	15.7
	Ensemble	14.8	11.9	12.7	13.5	14.3	14.2	15.9	15.1	15.1	14.4	15.2
Sharpe ratio	Full	0.73	0.43	0.68	0.21	0.85	0.91	0.78	0.79	0.74	0.88	0.88
	Size	0.7	1.23	0.66	0.76	1.07	1.41	1.19	1.25	0.97	1.29	1.25
	Ensemble	0.79	1.07	0.87	0.9	1.03	1.26	1.09	1.04	1.06	1.19	1.15
Max. drawdown (%)	Full	44.5	45.2	27.1	103.6	44.1	32.3	25.9	49.1	43.6	35.7	35.3
	Size	58.3	15.7	75.2	35.2	22.1	18.8	23.2	31.3	57.9	21.2	23.8
	Ensemble	44.9	45.2	20.0	20.3	26.3	22.2	25.8	33.5	32.2	23.7	26.6
Max. 1M loss (%)	Full	16.4	12.8	13.7	19.2	11.3	12.5	16.2	16.3	13.9	18.4	15.9
	Size	14.5	11.4	19.8	23.5	15.7	15.2	23.0	20.0	14.2	21.2	20.9
	Ensemble	16.7	9.5	16.6	9.8	15.3	13.5	16.5	14.6	10.1	16.2	13.6
Monthly one-way turnover (ann. %)	Full	128.4	116.3	112.2	148.3	133.5	132.7	133.9	135.2	135.4	139.4	136.0
	Size	124.4	119.2	124.0	142.1	136.1	137.9	136.2	134.1	129.9	137.6	136.1
	Ensemble	127.9	151.4	119.8	149.0	138.1	137.2	138.1	137.4	134.9	142.0	139.2
FF6 α (ann. %)	Full	8.3	3.8	8.8	2.9	9.5	10.3	9.6	9.6	9.0	10.1	9.9
	Size	9.5	15.0	7.9	10.8	14.7	20.8	18.5	16.7	11.1	19.4	18.8
	Ensemble	9.8	12.4	10.1	11.8	13.4	16.5	16.9	14.7	13.6	15.6	15.9
FF6 t-stat	Full	3.56	1.69	3.69	1.05	3.73	4.13	4.07	3.73	3.25	4.55	4.01
	Size	3.27	7.09	3.95	3.94	6.49	7.27	6.66	6.77	4.38	6.44	6.52
	Ensemble	3.76	5.72	5.1	5.37	5.32	6.46	5.93	5.34	4.52	6.6	5.83
FF6 break-even cost (bps)	Full	26.9	13.8	32.6	8.1	29.7	32.3	30.0	29.7	27.6	30.2	30.3
	Size	31.7	52.4	26.5	31.8	44.9	62.9	56.5	51.9	35.7	58.7	57.5
	Ensemble	31.8	34.2	35.3	33.0	40.4	50.1	50.9	44.5	42.1	45.7	47.7
HXZ break-even cost (bps)	Full	23.8	13.7	33.7	6.2	26.6	30.1	27.6	27.3	28.7	28.4	28.0
	Size	28.6	49.2	23.0	28.3	40.8	58.0	51.7	48.8	33.5	53.2	54.3
	Ensemble	27.9	31.8	33.8	27.1	36.5	48.2	46.7	40.1	43.6	40.6	43.1

Table 4.15: Out-of-sample performance of machine learning portfolios under different training regimes in the middle 40% of stocks

This table presents VW top-minus-bottom portfolio statistics for machine learning models trained using three approaches: Full, Size, and Ensemble. Results are presented within the middle stock universe, i.e., the middle 40% of stocks. The Full model follows the standard approach using all available stocks, whereas the Size model trains three separate models for large, mid, and small stocks and then concatenates the predictions to form portfolios. The Ensemble model is the average of the return predictions from the Full model and Size model. The table presents the performance statistics of these portfolios, including the annualized mean, standard deviation, Sharpe ratios, maximum drawdown, maximum one-month loss, average monthly one-way turnover, annualized FF6 alpha and t-statistic, and break-even transaction costs under the FF6 and HXZ risk models. The OOS period is from January 1987 to December 2021.

Metric	Model	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Portfolio return (ann. %)	Full	23.3	16.4	17.5	15.1	29.7	28.7	28.9	30.1	29.1	31.5	30.8
	Size	24.0	31.2	13.8	23.5	37.2	37.6	37.4	37.5	35.8	36.7	39.1
	Ensemble	25.8	30.8	18.0	23.1	37.4	36.3	36.9	36.2	34.9	37.8	38.5
Volatility (ann. %)	Full	18.0	12.2	14.0	15.8	14.7	14.4	15.6	15.4	15.0	15.7	15.4
	Size	20.9	17.5	18.0	18.7	17.8	18.8	18.7	18.4	18.8	19.6	18.7
	Ensemble	20.3	15.2	17.3	13.5	16.9	17.1	18.0	17.4	16.8	18.3	18.0
Sharpe ratio	Full	1.29	1.34	1.25	0.96	2.01	2.0	1.85	1.95	1.94	2.01	2.00
	Size	1.15	1.78	0.77	1.25	2.09	2.0	2.0	2.04	1.91	1.87	2.09
	Ensemble	1.27	2.04	1.04	1.71	2.22	2.13	2.05	2.07	2.07	2.07	2.13
Max. drawdown (%)	Full	74.8	24.4	28.4	51.2	34.1	31.6	43.1	40.9	40.3	35.1	35.9
	Size	80.4	48.5	58.9	62.8	35.1	31.5	42.0	29.6	38.3	36.8	31.6
	Ensemble	78.1	28.9	45.9	35.6	32.9	30.7	41.6	36.7	36.7	36.9	32.2
Max. 1M loss (%)	Full	27.7	16.2	11.1	16.0	17.8	15.2	20.5	16.9	20.9	21.7	16.7
	Size	31.1	23.4	28.3	38.9	19.5	21.9	23.3	19.7	21.1	20.2	21.8
	Ensemble	28.1	14.9	24.3	16.9	19.8	20.7	20.2	19.6	21.2	21.9	20.4
Monthly one-way turnover (ann. %)	Full	119.2	116.3	104.9	144.3	131.0	131.6	131.6	131.1	131.1	135.9	132.5
	Size	121.0	140.1	113.1	134.7	130.9	130.0	130.5	130.3	128.5	129.5	130.5
	Ensemble	121.5	147.9	108.7	140.5	133.2	132.9	132.5	132.3	131.6	136.1	134.5
FF6 α (ann. %)	Full	21.2	15.5	17.6	12.3	27.7	26.6	26.6	27.5	26.5	28.6	28.2
	Size	22.7	30.6	12.2	22.5	35.4	35.5	35.3	35.7	33.7	34.9	37.0
	Ensemble	24.0	30.0	17.2	21.3	35.4	34.4	34.7	34.6	32.7	35.9	36.1
FF6 t-stat	Full	5.38	5.8	6.04	4.01	8.23	8.84	7.8	7.84	7.99	8.17	8.15
	Size	4.74	6.93	3.17	5.54	7.83	8.05	7.61	8.14	7.49	7.46	8.13
	Ensemble	5.18	8.18	4.86	7.51	8.53	8.53	8.1	7.87	8.09	8.01	8.24
FF6 break-even cost (bps)	Full	74.1	55.5	69.7	35.5	88.1	84.2	84.2	87.4	84.2	87.8	88.8
	Size	78.1	91.0	44.8	69.5	112.7	113.9	112.8	114.1	109.2	112.2	118.2
	Ensemble	82.4	84.6	66.0	63.0	110.6	107.9	109.0	108.9	103.5	110.0	111.8
HXZ break-even cost (bps)	Full	69.4	53.5	64.9	34.8	85.2	79.9	81.8	85.0	83.8	84.1	87.1
	Size	75.9	88.1	49.5	66.8	111.1	111.2	111.1	112.5	108.5	108.5	116.9
	Ensemble	79.0	82.3	65.9	61.3	107.9	105.1	106.4	106.6	103.1	105.4	109.0

Table 4.16: Out-of-sample performance of machine learning portfolios under different training regimes in bottom 30% of stocks

This table presents VW top-minus-bottom portfolio statistics for machine learning models trained using three approaches: Full, Size, and Ensemble. Results are presented within the smallest stock universe, i.e., the smallest 30% of stocks. The Full model follows the standard approach using all available stocks, while the Size model trains three separate models for large, mid, and small stocks and then concatenates the predictions to form portfolios. The Ensemble model is the average of the return predictions from the Full model and Size model. The table presents the performance statistics of these portfolios, including the annualized mean, standard deviation, Sharpe ratios, maximum drawdown, maximum one-month loss, average monthly one-way turnover, annualized FF6 alpha and t-statistic, and break-even transaction costs under the FF6 and HXZ risk models. The OOS period is from January 1987 to December 2021.

Metric	Model	OLS	ENET	RF	GBRT	NN1	NN2	NN3	NN4	NN5	ENS	ENSNN
Portfolio return (ann. %)	Full	33.6	29.4	27.1	38.8	54.0	51.0	52.2	52.1	53.7	57.9	55.7
	Size	32.4	45.3	10.2	31.6	63.6	64.7	64.8	63.4	60.8	62.6	66.8
	Ensemble	35.9	49.2	19.4	44.3	64.4	60.9	62.6	62.5	60.0	66.3	66.4
Volatility (ann. %)	Full	21.6	14.6	16.3	26.6	15.9	15.2	15.1	16.2	17.2	18.3	15.8
	Size	21.1	20.1	13.9	19.2	18.5	19.6	19.6	19.8	20.5	19.1	20.4
	Ensemble	23.0	16.9	14.5	25.7	18.5	17.6	18.6	17.9	19.2	19.4	19.4
Sharpe ratio	Full	1.56	2.02	1.66	1.46	3.4	3.35	3.45	3.22	3.12	3.16	3.52
	Size	1.53	2.25	0.74	1.65	3.44	3.3	3.31	3.21	2.96	3.28	3.28
	Ensemble	1.56	2.9	1.34	1.72	3.48	3.47	3.37	3.49	3.13	3.41	3.43
Max. drawdown (%)	Full	115.1	19.8	16.5	48.4	16.8	17.4	10.3	13.5	17.0	35.0	14.7
	Size	92.3	38.7	40.5	37.6	19.1	13.7	17.8	14.1	18.4	19.8	15.0
	Ensemble	104.8	19.5	21.2	25.5	15.9	13.3	16.4	18.9	14.3	14.9	13.3
Max. 1M loss (%)	Full	38.0	12.0	16.5	15.9	15.1	14.0	10.3	11.6	17.0	26.7	14.7
	Size	36.0	16.0	16.6	15.4	12.4	9.6	9.7	10.7	12.9	9.1	12.0
	Ensemble	37.3	10.3	12.4	14.4	11.9	9.1	13.9	17.2	12.7	12.7	10.8
Monthly one-way turnover (ann. %)	Full	118.0	118.1	106.4	149.5	142.2	142.0	142.1	141.1	140.5	145.9	143.4
	Size	129.4	163.3	123.3	146.2	139.5	139.4	139.3	139.4	139.3	143.7	139.9
	Ensemble	125.9	161.4	121.0	151.7	142.8	141.8	142.1	141.3	140.4	147.5	144.1
FF6 α (ann. %)	Full	32.5	30.4	26.6	35.1	52.5	49.8	50.9	50.5	51.7	55.2	53.8
	Size	31.3	43.9	10.0	30.9	61.9	62.5	61.6	61.4	59.3	61.0	64.8
	Ensemble	34.9	48.4	18.2	40.9	62.3	59.3	60.1	60.7	57.2	64.1	64.4
FF6 t-stat	Full	6.93	8.42	7.85	5.78	13.09	13.62	13.77	12.92	11.49	11.47	13.57
	Size	6.29	9.82	3.89	8.03	12.64	12.37	12.54	12.12	11.06	12.07	12.11
	Ensemble	6.44	12.26	6.02	6.99	12.77	13.91	12.81	13.58	11.87	12.32	13.1
FF6 break-even cost (bps)	Full	114.6	107.4	104.3	97.9	153.9	146.0	149.3	149.0	153.3	157.7	156.3
	Size	100.9	112.1	33.9	88.0	184.8	186.9	184.4	183.5	177.5	176.9	193.0
	Ensemble	115.5	125.0	62.5	112.3	181.7	174.1	176.3	179.0	169.7	181.2	186.1
HXZ break-even cost (bps)	Full	108.5	103.0	103.9	96.8	147.9	138.9	145.8	143.8	147.5	151.9	151.9
	Size	97.5	108.0	30.6	87.2	177.2	180.0	179.0	176.5	168.0	170.2	184.3
	Ensemble	112.1	121.8	62.3	110.1	175.4	168.8	170.8	173.1	163.8	176.3	180.4

Chapter 5

Conclusions & future research

Through this thesis, I have challenged commonly held beliefs in empirical finance, presenting modern takes on traditional and more innovative models. Specifically, I explored three distinct topics in empirical finance:

1. Asynchronicity in financial time series
2. The death of the S&P index effect
3. Biases and overfitting when training machine learning models for return prediction

Although each chapter is distinct, a common thread of comparing traditional and innovative models persists. I have emphasized the need for continuous innovation in empirical finance research by comparing, contrasting, and challenging commonly held beliefs.

5.1 Tradition and innovation

5.1.1 Asynchronicity in financial time series

Asynchronicity is a pervasive feature of financial time series, directly impacting inference in empirical models that assume contemporaneous measurement of time-series observations. This thesis has demonstrated the value of using DTW, a non-parametric, unsupervised machine learning technique, to measure and correct for asynchronicity between financial time series. This innovative approach has allowed for new insights into market behaviors, such as the CAPM beta anomaly and the analysis of intraday price leadership between liquid futures contracts. The DTW

technique has scope for applications in other areas of financial economics where asynchronicity drives misestimation and inference in modeling.

Future research on asynchronicity could further explore the potential of DTW in applications such as disentangling systematic and idiosyncratic risk changes when stocks are added to an index or measuring intraday betas to study diurnal patterns in intraday stock returns. Moreover, a formal price discovery measure using DTW can be developed.

5.1.2 The S&P index effect

By examining a full sample of S&P 500, S&P 400, and S&P 600 index changes, this thesis has shown that the S&P index effect has not disappeared but rather migrated. The index effect is still present for stocks added to the S&P U.S. indexes from outside the combined universe of S&P 1500 stocks. This finding has implications for passive investment managers, allowing them to adjust their portfolios in response to changing market dynamics. In addition, this research has demonstrated the presence of informed traders in the options market, who trade ahead of index announcements and attempt to profit from such changes.

Future research could expand upon these results by applying the same analysis to global indexes, such as the MSCI World, MSCI Small Cap, and MSCI Emerging Markets indexes, to further understand the implications of index change procedures on underlying stock return behavior. With the continued growth of passive investing and the increasing importance of index vendors such as S&P and MSCI, understanding the implications of index vendor policies on the functioning of financial markets and ensuring that unintended consequences from such mechanical behaviors are limited will be crucial.

5.1.3 Training machine learning models in finance

This thesis has explored the role of data and customization in training machine learning models for stock return prediction, demonstrating that splitting the current training approach into group–specific sub-samples can achieve superior returns compared with training on the full sample. This finding counters the intuition that "more data are better" when training machine learning models and highlights the potential performance gains that can be achieved over the current approaches in the literature. In particular, the current approach of predicting raw excess returns can introduce various biases into the machine learning training process, producing sub-optimal predictions for a large portion of the stock universe of interest. Superior

machine learning model performance for predicting cross-sectional returns can be achieved by applying simple regularization to raw excess returns.

Future research in this area can focus on the relationship between the choice of training data and machine learning return predictions and on the interpretability of machine learning models in finance. Interpretability is critical to the success and longevity of machine learning models in finance, particularly for use by regulated entities such as banks. Further, I trained three independent size-specific models in this study and simply concatenated the predictions. Future research could explore linking mechanisms between these models, such as using the previous month's performance of the machine learning model trained on large stocks as an input feature to the small stock prediction model.

An important area of future research is exploring the optimal size of the data used to predict stock returns in the formation of trading strategies. The simplest approach is to simultaneously predict the entire cross-section using a single parsimonious model. However, I have shown how in a machine learning framework, such an approach can lead to sub-optimal outcomes. At the other end of the spectrum is individually fitting models for each stock to predict returns, and then concatenating all the predictions into a trading strategy. Such an approach is both computationally expensive and cannot easily take advantage of the known cross-sectional factor structure in stock returns. Thus, understanding what the ideal middle ground between these two scenarios for training machine learning models to price the cross-section is is an important future research area.

5.2 Implications and applications

This thesis contributes to the field of financial economics and asset pricing by addressing the gaps in the understanding of asynchronicity in financial time series, the evolving nature of the S&P index effect, and the application of machine learning models in asset pricing. By challenging conventional beliefs and presenting novel approaches, this research helps enhance the accuracy and effectiveness of financial models. In this study, the development and application of innovative techniques, such as DTW, has led to the discovery of new insights into market behaviors, and accordingly to more informed decision-making processes for investors, analysts, and policymakers.

The implications of this research extend beyond academia and to real-world applications in the financial industry. The insights gained from analyzing asynchronicity can lead to more accurate risk measurement and better portfolio

risk management. The findings on the evolution and migration of the S&P index effect offer valuable information to passive asset managers, enabling them to adjust their portfolios in response to changing market dynamics. Additionally, exploring machine learning models in finance provides a foundation for more sophisticated and precise prediction models that can help guide investment decisions and risk management. By acknowledging the limitations of traditional models and embracing innovative approaches, this thesis ultimately helps develop more accurate models of financial markets.

5.3 Future research

There are several areas where future research can further build upon the findings of this thesis:

1. Developing more advanced methods for measuring asynchronicity that address the limitations of DTW and exploring other potential applications of DTW.
2. Measuring the recent behavior of the index effect in other international indexes, such as MSCI World, MSCI Small Cap, and MSCI Emerging Markets
3. Investigating the selection method of training data for machine learning models in finance, focusing on understanding the relationship between the choice of training data and subsequent machine learning return predictions.
4. Exploring how to create machine learning models that hedge poor performance of traditional asset pricing models.

Bibliography

Ambachtsheer, K. P. (1974), Profit potential in an "almost efficient" market, *Journal of Portfolio Management* **1**(1), 84–87.

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

An, B.-J., Ang, A., Bali, T. G. and Cakici, N. (2014), The joint cross section of stocks and options, *Journal of Finance* **69**(5), 2279–2337.

Anadu, K., Kruttli, M., McCabe, P. and Osambela, E. (2020), The shift from active to passive investing: Risks to financial stability?, *Financial Analysts Journal* **76**(4), 23–39.

Antonakakis, N., Floros, C. and Kizys, R. (2016), Dynamic spillover effects in futures markets: UK and US evidence, *International Review of Financial Analysis* **48**, 406–418.

Arnott, R., Brightman, C., Kalesnik, V. and Wu, L. (2022), The avoidable costs of index rebalancing, *Working paper* .

Audrino, F. and Bühlmann, P. (2004), Synchronizing multivariate financial time series, *Journal of Risk* **6**(2), 81–106.

Augustin, P. and Subrahmanyam, M. G. (2020), Informed options trading before corporate events, *Annual Review of Financial Economics* **12**(1), 327–355.

Avramov, D., Cheng, S. and Metzker, L. (2022), Machine learning vs. economic restrictions: Evidence from stock return predictability, *Management Science* **69**(5), 2587–2619.

Azevedo, V. and Hoegner, C. (2023), Enhancing stock market anomalies with machine learning, *Review of Quantitative Finance and Accounting* **60**, 195–230.

Azevedo, V., Kaiser, G. S., Kaiser, S. and Muller, S. (2023), Stock market anomalies and machine learning across the globe, *Journal of Asset Management* **24**, 419–441.

Bali, T. G., Brown, S., Murray, S. and Tang, Y. (2017), A lottery-demand-based explanation of the beta anomaly, *Journal of Financial and Quantitative Analysis* **52**(6), 2369–2397.

Bali, T. G., Cakici, N. and Whitelaw, R. F. (2011), Maxing out: Stocks as lotteries and the cross-section of expected returns, *Journal of Financial Economics* **99**(2), 427–446.

Barberis, N., Shleifer, A. and Wurgler, J. (2005), Comovement, *Journal of Financial Economics* **75**(2), 283–317.

Becker-Blease, J. R. and Paul, D. L. (2010), Does inclusion in a smaller S&P index create value?, *Financial Review* **45**(2), 307–330.

Ben-David, I., Franzoni, F. and Moussawi, R. (2018), Do ETFs increase volatility?, *Journal of Finance* **73**(6), 2471–2535.

Bender, J., Nagori, R. and Tank, M. (2019), The past, present, and future of the index effect, *Journal of Beta Investment Strategies* **10**(3), 15–37.

Beneish, M. D., Lee, C. M. C. and Tarpley, R. L. (2001), Contextual fundamental analysis through the prediction of extreme returns, *Review of Accounting Studies* **6**(2/3), 165–189.

Bennett, B., Stulz, R. M. and Wang, Z. (2020), Does joining the S&P 500 index hurt firms?, *Working paper* .

Biktimirov, E. N. and Xu, Y. (2019), Asymmetric stock price and investor awareness reactions to changes in the Nasdaq 100 index, *Journal of Asset Management* **20**(2), 134–145.

Black, F., Jensen, M. and Scholes, M. (1972), The capital asset pricing model: Some empirical tests, *Journal of Financial Economics* **128**(1), 79–121.

Blitz, D. C. and van Vliet, P. (2007), The volatility effect, *Journal of Portfolio Management* **34**(1), 102–113.

Blitz, D. and Hanauer, M. X. (2020), Settling the size matter, *Journal of Portfolio Management* **47**(2), 99–112.

Blitz, D., Hanauer, M. X., Hoogteijling, T. and Howard, C. (2023), The term structure of machine learning alpha, *Journal of Financial Data Science* (forthcoming).

Blitz, D., Hoogteijling, T., Lohre, H. and Messow, P. (2023), How can machine learning advance quantitative asset management?, *Journal of Portfolio Management* (forthcoming).

Blitz, D., van Vliet, P. and Baltussen, G. (2019), The volatility effect revisited, *Journal of Portfolio Management* **46**(2), 45–63.

Burns, P., Engle, R. F. and Mezrich, J. J. (1998), Correlations and volatilities of asynchronous data, *Journal of Derivatives* **5**(4), 7–18.

Cai, J. and Houge, T. (2008), Long-term impact of Russell 2000 index rebalancing, *Financial Analysts Journal* **64**(4), 76–91.

Cakici, N., Fieberg, C., Metko, D. and Zaremba, A. (2023), Machine learning goes global: Cross-sectional return predictability in international stock markets, *Journal of Economic Dynamics and Control* **155**, 104725.

Carhart, M. M. (1997), On persistence in mutual fund performance, *Journal of Finance* **52**(1), 57–82.

Chakrabarti, R., Huang, W., Jayaraman, N. and Lee, J. (2005), Price and volume effects of changes in MSCI indices–nature and causes, *Journal of Banking & Finance* **29**(5), 1237–1264.

Chen, A. Y. and Zimmermann, T. (2022), Open source cross-sectional asset pricing, *Critical Finance Review* **11**(2), 207–264.

Chen, H., Noronha, G. and Singal, V. (2004), The price response to S&P 500 index additions and deletions: Evidence of asymmetry and a new explanation, *Journal of Finance* **59**(4), 1901–1930.

Chen, L., Pelger, M. and Zhu, J. (2023), Deep learning in asset pricing, *Management Science* (forthcoming).

Chen, Y., Koutsantony, C., Truong, C. and Veeraraghavan, M. (2013), Stock price response to S&P 500 index inclusions: Do options listings and options trading volume matter?, *Journal of International Financial Markets, Institutions and Money* **23**, 379–401.

Choi, D., Jiang, W. and Zhang, C. (2019), Alpha go everywhere: Machine learning and international stock returns, *Working paper* .

Chu, G., Goodell, J. W., Li, X. and Zhang, Y. (2021), Long-term impacts of index reconstitutions: Evidence from the CSI 300 additions and deletions, *Pacific-Basin Finance Journal* **69**, 101651.

Coates, J. C. (2018), The future of corporate governance part I: The problem of twelve, *Working paper* .

Cohen, K. J., Hawawini, G. A., Maier, S. F., Schwartz, R. A. and Whitcomb, D. K. (1983), Friction in the trading process and the estimation of systematic risk, *Journal of Financial Economics* **12**(2), 263–278.

Cong, L. W., Feng, G., He, J. and He, X. (2022), Asset pricing with panel tree under global split criteria, *Working paper* .

Cong, L. W., Feng, G., He, J. and Li, J. (2022), Uncommon factors for Bayesian asset clusters, *Working paper* .

Conrad, J., Dittmar, R. F. and Ghysels, E. (2013), Ex ante skewness and expected stock returns, *Journal of Finance* **68**(1), 85–124.

Da, Z. and Shive, S. (2018), Exchange traded funds and asset return correlations, *European Financial Management* **24**(1), 136–168.

DeMiguel, V., Nogales, F. J. and Uppal, R. (2014), Stock return serial dependence and out-of-sample portfolio performance, *Review of Financial Studies* **27**(4), 1031–1073.

Dessain, J. (2022), Machine learning models predicting returns: Why most popular performance metrics are misleading and proposal for an efficient metric, *Expert Systems with Applications* **199**, 116970.

Dhillon, U. and Johnson, H. (1991), Changes in the Standard and Poor's 500 list, *Journal of Business* **64**(1), 75–85.

Diebold, F. and Mariano, R. (1995), Comparing predictive accuracy, *Journal of Business & Economic Statistics* **13**(3), 253–63.

Dietterich, T. G. (2000), Ensemble methods in machine learning, *in* 'Multiple Classifier Systems', Vol. 1857, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–15.

Dimson, E. (1979), Risk measurement when shares are subject to infrequent trading, *Journal of Financial Economics* **7**(2), 197–226.

Dobrev, D. and Schaumburg, E. (2016), High-frequency cross-market trading: Model free measurement and applications, *Working Paper* .

Docking, D. S. and Dowen, R. J. (2006), Evidence on stock price effects associated with changes in the S&P 600 SmallCap index, *Quarterly Journal of Business and Economics* **45**(1/2), 89–114.

Doshi, H., Elkamhi, R. and Simutin, M. (2015), Managerial activeness and mutual fund performance, *Review of Asset Pricing Studies* **5**(2), 156–184.

Drobetz, W. and Otto, T. (2021), Empirical asset pricing via machine learning: Evidence from the European stock market, *Journal of Asset Management* **22**(7), 507–538.

Eun, C. S. and Shim, S. (1989), International transmission of stock market movements, *Journal of Financial and Quantitative Analysis* **24**(2), 241–256.

Fama, E. F. and French, K. R. (1992), The cross-section of expected stock returns, *Journal of Finance* **47**(2), 427–465.

Fama, E. F. and French, K. R. (1993), Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* **33**(1), 3–56.

Fama, E. F. and French, K. R. (2015), A five-factor asset pricing model, *Journal of Financial Economics* **116**(1), 1–22.

Fama, E. F. and French, K. R. (2018), Choosing factors, *Journal of Financial Economics* **128**(2), 234–252.

Fama, E. F. and MacBeth, J. D. (1973), Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy* **81**(3), 607–636.

Feng, G., He, J. and Polson, N. G. (2018), Deep learning for predicting asset returns, *Working paper* .

Feng, G., Polson, N. G. and Xu, J. (2018), Deep learning in characteristics-sorted factor models, *Working paper* .

Fernandes, M. and Mergulhão, J. (2016), Anticipatory effects in the FTSE 100 index revisions, *Journal of Empirical Finance* **37**, 79–90.

Franses, P. H. and Wiemann, T. (2020), Intertemporal similarity of economic time series: An application of dynamic time warping, *Computational Economics* **56**(1), 59–75.

Frazzini, A. and Pedersen, L. H. (2014), Betting against beta, *Journal of Financial Economics* **111**(1), 1–25.

Freyberger, J., Neuhierl, A. and Weber, M. (2020), Dissecting characteristics nonparametrically, *The Review of Financial Studies* **33**(5), 2326–2377.

Fu, X., Arisoy, Y. E., Shackleton, M. B. and Umutlu, M. (2016), Option-implied volatility measures and stock return predictability, *Journal of Derivatives* **24**(1), 58–78.

Glosten, L., Nallareddy, S. and Zou, Y. (2021), ETF activity and informational efficiency of underlying securities, *Management Science* **67**(1), 22–47.

Gonzalo, J. and Granger, C. (1995), Estimation of common long-memory components in cointegrated systems, *Journal of Business & Economic Statistics* **13**(1), 27–35.

Gowri Shankar, S. and Miller, J. M. (2006), Market reaction to changes in the S&P SmallCap 600 index, *Financial Review* **41**(3), 339–360.

Grinold, R. C. (1989), The fundamental law of active management, *Journal of Portfolio Management* **15**(3), 30–37.

Gu, S., Kelly, B. and Xiu, D. (2020), Empirical asset pricing via machine learning, *The Review of Financial Studies* **33**(5), 2223–2273.

Hanauer, M. X. and Kalsbach, T. (2023), Machine learning and the cross-section of emerging market stock returns, *Emerging Markets Review* **55**, 101022.

Hansen, L. and Salamon, P. (1990), Neural network ensembles, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(10), 993–1001.

Harris, L. and Gurel, E. (1986), Price and volume effects associated with changes in the S&P 500 list: New evidence for the existence of price pressures, *Journal of Finance* **41**(4), 815–829.

Harvey, C. R. and Liu, Y. (2021), Lucky factors, *Journal of Financial Economics* **141**(2), 413–435.

Harvey, C. R. and Siddique, A. (2000), Conditional skewness in asset pricing tests, *Journal of Finance* **55**(3), 1263–1295.

Hasbrouck, J. (1995), One security, many markets: Determining the contributions to price discovery, *Journal of Finance* **50**(4), 1175–1199.

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

Heaton, J. B., Polson, N. G. and Witte, J. H. (2017), Deep learning for finance: Deep portfolios, *Applied Stochastic Models in Business and Industry* **33**(1), 3–12.

Hollstein, F. and Simen, C. W. (2021), The index effect: Evidence from the option market, *Working paper* .

Hou, K., Xue, C. and Zhang, L. (2015), Digesting anomalies: An investment approach, *Review of Financial Studies* **28**(3), 650–705.

Israel, R., Kelly, B. and Moskowitz, T. (2020), Can machines "learn" finance?, *Journal of Investment Management* **18**(2).

Ito, K. and Sakemoto, R. (2020), Direct estimation of lead–lag relationships using multinomial dynamic time warping, *Asia-Pacific Financial Markets* **27**(3), 325–342.

Jain, P. C. (1987), The effect on stock price of inclusion in or exclusion from the S&P 500, *Financial Analysts Journal* **43**(1), 58–65.

Jegadeesh, N. (1990), Evidence of predictable behavior of security returns, *Journal of Finance* **45**(3), 881–898.

Jegadeesh, N. and Titman, S. (1993), Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* **48**(1), 65–91.

Jeong, Y.-S., Jeong, M. K. and Omitaomu, O. A. (2011), Weighted dynamic time warping for time series classification, *Pattern Recognition* **44**(9), 2231–2240. Computer Analysis of Images and Patterns.

Kamal, R., Lawrence, E. R., McCabe, G. and Prakash, A. J. (2012), Additions to S&P 500 index: Not so informative any more, *Managerial Finance* **38**(4), 380–402.

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

Kelly, B. T., Pruitt, S. and Su, Y. (2019), Characteristics are covariances: A unified model of risk and return, *Journal of Financial Economics* **134**(3), 501–524.

Keogh, E. and Pazzani, M. (2002), Derivative dynamic time warping, *First SIAM International Conference on Data Mining* **1**, 1–11.

Kim, C. W., Li, X. and Perry, T. T. (2017), Adaptation of the S&P 500 index effect, *Journal of Index Investing* **8**(1), 29–36.

Kozak, S., Nagel, S. and Santosh, S. (2020), Shrinking the cross-section, *Journal of Financial Economics* **135**(2), 271–292.

Lahreche, A. and Boucheham, B. (2021), A comparison study of dynamic time warping's variants for time series classification, *International Journal of Informatics and Applied Mathematics* **4**(1), 56–71.

Lakonishok, J. and Shapiro, A. C. (1986), Systematic risk, total risk and size as determinants of stock market returns, *Journal of Banking & Finance* **10**(1), 115–132.

Lalwani, V. and Meshram, V. V. (2022), The cross-section of Indian stock returns: Evidence using machine learning, *Applied Economics* **54**(16), 1814–1828.

Laughlin, G., Aguirre, A. and Grundfest, J. (2014), Information transmission between financial markets in Chicago and New York, *Financial Review* **49**(2), 283–312.

Ledoit, O. and Wolf, M. (2008), Robust performance hypothesis testing with the Sharpe ratio, *Journal of Empirical Finance* **15**(5), 850–859.

Lehmann, B. N. (1990), Residual risk revisited, *Journal of Econometrics* **45**(1-2), 71–97.

Leippold, M., Wang, Q. and Zhou, W. (2022), Machine learning in the Chinese stock market, *Journal of Financial Economics* **145**(2), 64–82.

Leung, E., Lohre, H., Mischlich, D., Shea, Y. and Stroh, M. (2021), The promises and pitfalls of machine learning for predicting stock returns, *Journal of Financial Data Science* **3**(2), 21–50.

Li, M., Yin, X. and Zhao, J. (2020), Does program trading contribute to excess comovement of stock returns?, *Journal of Empirical Finance* **59**, 257–277.

Li, Y., Simon, Z. and Turkington, D. (2022), Investable and interpretable machine learning for equities, *Journal of Financial Data Science* **4**(1), 54–74.

Liu, J., Stambaugh, R. F. and Yuan, Y. (2018), Absolving beta of volatility's effects, *Journal of Financial Economics* **128**(1), 1–15.

Liu, Q., Tao, Z., Tse, Y. and Wang, C. (2022), Stock market prediction with deep learning: The case of China, *Finance Research Letters* **46**, 102209.

Lo, A. W. and MacKinlay, A. C. (1990), An econometric analysis of nonsynchronous trading, *Journal of Econometrics* **45**(1), 181–211.

Lu, L., Shin, Y., Su, Y. and Karniakdakis, G. E. (2020), Dying ReLU and initialization: Theory and numerical examples, *Communications in Computational Physics* **28**(5), 1671–1706.

Lynch, A. W. and Mendenhall, R. R. (1997), New evidence on stock price effects associated with changes in the S&P 500 index, *Journal of Business* **70**(3), 351–383.

Malceniece, L., Malcenieks, K. and Putniņš, T. J. (2019), High frequency trading and comovement in financial markets, *Journal of Financial Economics* **134**(2), 381–399.

Marciniak, M. (2010), Information effects of announced stock index additions: Evidence from S&P 400, *Journal of Economics and Finance* **36**(4), 822–849.

Martens, M. and Poon, S.-H. (2001), Returns synchronization and daily correlation dynamics between international stock markets, *Journal of Banking & Finance* **25**(10), 1805–1827.

Mase, B. (2007), The impact of changes in the FTSE 100 index, *Financial Review* **42**(3), 461–484.

Masse, I., Hanrahan, R., Kushner, J. and Martinello, F. (2000), The effect of additions to or deletions from the TSE 300 index on Canadian share prices, *Canadian Journal of Economics* **33**(2), 341–359.

Moritz, B. and Zimmermann, T. (2016), Tree-based conditional portfolio sorts: The relation between past and future stock returns, *Working paper* .

Novy-Marx, R. and Velikov, M. (2022), Betting against betting against beta, *Journal of Financial Economics* **143**(1), 80–106.

Ozturk, S. R., van der Wel, M. and van Dijk, D. (2017), Intraday price discovery in fragmented markets, *Journal of Financial Markets* **32**, 28–48.

Pan, J. and Poteshman, A. (2006), The information in option volume for future stock prices, *Review of Financial Studies* **19**, 871–908.

Park, A. and Wang, J. (2020), Did trading bots resurrect the CAPM?, *Working paper* .

Petajisto, A. (2011), The index premium and its hidden cost for index funds, *Journal of Empirical Finance* **18**(2), 271–288.

Piotroski, J. D. (2000), Value investing: The use of historical financial statement information to separate winners from losers, *Journal of Accounting Research* **38**, 1–41.

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

Qian, J. and Su, L. (2016), Shrinkage estimation of common breaks in panel data models via adaptive group fused lasso, *Journal of Econometrics* **191**(1), 86–109.

Rasekhschaffe, K. C. and Jones, R. C. (2019), Machine learning for stock selection, *Financial Analysts Journal* **75**(3), 70–88.

Rauterberg, G. V. and Verstein, A. (2013), Index theory: The law, promise and failure of financial indices, *Yale Journal on Regulation* **30**(1), 1–62.

Reinganum, M. R. (1981), Misspecification of capital asset pricing: Empirical anomalies based on earnings' yields and market values, *Journal of Financial Economics* **9**(1), 19–46.

Rua, A. and Nunes, L. C. (2009), International comovement of stock market returns: A wavelet analysis, *Journal of Empirical Finance* **16**(4), 632–639.

Rubesam, A. (2022), Machine learning portfolios with equal risk contributions: Evidence from the Brazilian market, *Emerging Markets Review* **51**, 100891.

Sakoe, H. and Chiba, S. (1978), Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**(1), 43–49.

Sammon, M. (2022), Passive ownership and price informativeness, *Working paper* .

Scherer, B. (2013), Synchronize your data or get out of step with your risks, *Journal of Derivatives* **20**, 75–84.

Scholes, M. S. and Williams, J. (1977), Estimating betas from nonsynchronous data, *Journal of Financial Economics* **5**(3), 309–327.

SEC (2020), SEC charges index manager and friend with insider trading [Press Release]. https://www.sec.gov/news/press-release/2020-217, last accessed on 2022-09-04.

Shleifer, A. (1986), Do demand curves for stocks slope down?, *Journal of Finance* **41**(3), 579–590.

Shumway, T. (1997), The delisting bias in CRSP data, *Journal of Finance* **52**(1), 327–340.

Shumway, T. and Warther, V. A. (1999), The delisting bias in CRSP's NASDAQ data and its implications for the size effect, *Journal of Finance* **54**(6), 2361–2379.

Sorensen, E. H., Hua, R. and Qian, E. (2005), Contextual fundamentals, models, and active management, *Journal of Portfolio Management* **32**(1), 23–36.

Sushko, V. and Turner, G. (2018), The implications of passive investing for securities markets, *BIS Quarterly Review* pp. 113–131.

Tobek, O. and Hronec, M. (2021), Does it pay to follow anomalies research? Machine learning approach with international evidence, *Journal of Financial Markets* **56**, 100588.

Vijh, A. M. and Wang, J. B. (2022), Negative returns on addition to the S&P 500 index and positive returns on deletion? New evidence on the attractiveness of S&P 500 versus S&P 400 indexes, *Financial Management* **51**(4), 1127–1164.

Weller, B. M. (2018), Does algorithmic trading reduce information acquisition?, *Review of Financial Studies* **31**(6), 2184–2226.

Wong, S. Y. K., Chan, J. S. K., Azizi, L. and Xu, R. Y. D. (2022), Time-varying neural network for stock return prediction, *Intelligent Systems in Accounting, Finance and Management* **29**(1), 3–18.

Yu, S., Webb, G. and Tandon, K. (2014), What happens when a stock is added to the NASDAQ-100 index? What doesn't happen?, *Managerial Finance* **41**(5), 480–506.

Yun, J. and Kim, T. S. (2010), The effect of changes in index constitution: Evidence from the Korean stock market, *International Review of Financial Analysis* **19**(4), 258–269.

Zhang, Z., Tang, P. and Duan, R. (2015), Dynamic time warping under pointwise shape context, *Information Sciences* **315**, 88–101.