Université de Neuchâtel
Faculté des sciences économiques

THREE ESSAYS ON PRICE FORMATION AND LIQUIDITY IN FINANCIAL MARKETS

PhD thesis submitted to the Faculty of Economics and Business

Institute of Financial Analysis

University of Neuchâtel

For the PhD degree in Financial Markets

by

Emanuele Guidotti

Approved by the dissertation committee:

Prof. Tim A. KROENCKE, University of Neuchâtel, thesis director

Prof. David ARDIA, HEC Montréal, Canada, thesis co-director

Prof. Florian WEIGERT, University of Neuchâtel, jury president

Prof. Christian WAGNER, WU Vienna University, Austria

Dr. Jean-Philippe BOUCHAUD, Capital Fund Management (CFM);
École Normale Supérieure, France; Imperial College London, England

Defended on 26 April 2023

Faculté des sciences économiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel

www.unine.ch/seco

Université de Neuchâtel
Faculté des sciences économiques

IMPRIMATUR POUR LA THÈSE

Three Essays on Price Formation and Liquidity in Financial Markets

Emanuele GUIDOTTI

UNIVERSITÉ DE NEUCHÂTEL

FACULTÉ DES SCIENCES ÉCONOMIQUES

La Faculté des sciences économiques,

sur le rapport des membres du jury

Prof. Tim A. KROENCKE, Université de Neuchâtel, co-directeur de thèse

Prof. David ARDIA, HEC Montréal, co-directeur de thèse

Prof. Florian WEIGERT, Université de Neuchâtel

Prof. Jean-Philippe BOUCHAUD, Ecole Normale Supérieure et Imperial College London

Prof. Christian WAGNER, WU Vienna University of Economics and Business

autorise l'impression de la présente thèse.

Neuchâtel, le 30 mai 2023

Le doyen

Valéry Bezençon

Faculté des sciences économiques
Avenue du 1er-Mars 26
CH-2000 Neuchâtel
Email: secretariat.seco@unine.ch
Tel: +41 32 718 15 00
www.unine.ch/seco

"Measure what can be measured, and make measurable what cannot be measured."

—Galileo Galilei

Acknowledgments

First and foremost, I would like to extend my heartfelt thanks to my PhD supervisor, Tim A. Kroencke, for his invaluable guidance, mentorship, and support throughout my research journey. His friendly and supportive approach has made my PhD experience enjoyable and memorable.

I am also indebted to David Ardia, for his generous hospitality during my visit to Montréal and for providing invaluable assistance on numerous research projects.

Special thanks to Stefano M. Iacus for his mentorship and support in preparing me for this journey. To Carlo Gallina, for entrusting me with the opportunity to gain a unique perspective on the financial industry. To Alfio Ferrara, for his enthusiastic collaboration on our machine-learning projects. To Nakahiro Yoshida, for exposing me to statistical research and making me part of the YUIMA team. I also want to thank the Japan Science and Technology Agency CREST for their support.

I am grateful to my dissertation committee, David Ardia, Jean-Philippe Bouchaud, Tim A. Kroencke, Christian Wagner, and Florian Weigert, for kindly accepting to review this manuscript and contributing their time and effort to this thesis.

To the University of Neuchâtel, for providing all the necessary infrastructure and support to carry out my research, including funding for research visits and conferences worldwide. In particular, I want to thank Kira Facchinetti for her outstanding administrative support and kindness. To the professors Peter Fiechter, Tim A. Kroencke, Carolina Salva, and Florian Weigert for maintaining a relaxed yet stimulating research environment within the Institute of Financial Analysis. To my PhD colleagues Albian, Arnaud, Axel, Christophe, Daniele, Fanol, Keven, Loïc, Louis, Nikolay, Polina, and Sidita for the friendly and collaborative atmosphere.

I would also like to express my appreciation to Martine, Sandrine, and Fred for welcoming me to Switzerland and for their hospitality.

Finally, to my family, who has always been there for me and has encouraged me to chase my dreams. And to Sara, for her unending support, encouragement, and love throughout this journey and beyond.

Abstract

This dissertation studies how microstructure variables affect asset prices and participate in the price formation process. The dissertation consists of three chapters, of which the first two are co-written with Tim A. Kroencke and David Ardia.

The first chapter derives an efficient estimator of the bid-ask spread from open, high, low, and close prices. The estimator is asymptotically unbiased and optimally combines the full set of price data to minimize the estimation variance. The estimator makes it possible to estimate bid-ask spreads for a variety of assets regardless of their trading frequency, in low and high frequency, for markets that do not report bid and ask data, when quotes are unreliable or expensive to obtain, or when trades cannot be reliably matched with quotes. We find that earlier methods overstate bid-ask spreads where they are expected to be the smallest and understate bid-ask spreads where they are expected to be the largest, and we illustrate how our method improves inference in a variety of applications.

The second chapter uses the estimator to examine liquidity premia in the U.S. stock market since 1927, where liquidity is measured by the bid-ask spread. We find that the liquidity premium accounts for 20% of the total equity premium. Liquidity is more strongly priced in January, but it is also significant in non-January months. Our results suggest that liquidity is priced across all stocks, although samples restricted to larger stocks may lack power and achieve lower significance. Stocks with higher spreads are associated with lower prices of illiquidity and higher illiquidity premia. We also establish a link between bid-ask spreads and Amihud's illiquidity measure. In summary, we find that bid-ask spreads represent an important dimension of liquidity and they have pervasive effects on asset prices.

In the last chapter, I propose a theory where prices are formed in a purely mechanical manner through trading. The theory consists of three fundamental propositions. First, the quantity exchanged in a trade equals the integrated density of the book between zero and the peak impact. Second, the peak impact relaxes to the permanent impact such that makers and takers earn zero profits in the transaction. Third, the asset price is determined by the accumulation of price impacts. The model is simple yet capable of replicating various patterns observed in financial data. This work presents a mechanical view of financial markets in which trading itself is the source of financial market fluctuations and suggests a new research path in which the age-old question of what drives prices is replaced with what motivates people to trade in the first place.

Keywords: Financial Markets, Price Formation, Liquidity, Bid-Ask Spread.

Contents

List of Tables

List of Figures

Introduction

The bid-ask spread is a key measure in finance. However, an accurate measure requires high-frequency data, which are typically unavailable for international markets, asset classes other than stocks, and historical data samples. Accordingly, most studies either limit the sample to the time periods and markets of common data coverage or use proxies estimated from low-frequency data only. Following the seminal work by Roll (1984), several approaches have been proposed to estimate bid-ask spreads from daily prices. Lesmond et al. (1999) introduce a model based on the frequency of zero returns. Hasbrouck (2009) proposes a Gibbs estimation of the Roll model. Corwin & Schultz (2012) develop a simple way to estimate bid-ask spreads from high and low prices. Abdi & Ranaldo (2017) derive an estimator that jointly considers high, low, and close prices. However, these estimators are biased when trading is infrequent and when the spread is small relative to volatility (see Corwin & Schultz, 2012; Abdi & Ranaldo, 2017; Nieto, 2018; Jahan-Parvar & Zikes, 2019).

The first chapter of this dissertation derives an efficient estimator of the bid-ask spread from open, high, low, and close prices. The estimator is asymptotically unbiased and optimally combines the full set of price data to minimize the estimation variance. We compare our estimator with the seminal Roll (1984) estimator, and with Corwin & Schultz (2012) and Abdi & Ranaldo (2017), which have been shown to generally deliver the most accurate estimates in several comparative studies (Corwin & Schultz, 2012; Holden & Jacobsen, 2014; Karnaukh et al., 2015; Abdi & Ranaldo, 2017; Johann & Theissen, 2017). In our simulation experiments, we compare the bias and variance achieved by the estimators using daily data and an estimation window ranging from one month to one year. We also run the comparison for simulations performed in high-frequency, where we use minute data and an estimation window of one hour. Our method produces the lowest bias and variance in all scenarios, indicating it is always the best choice regardless of the estimation window and the evaluation metric used by the researcher, both in low and high frequencies. Our empirical analysis uses the CRSP U.S. stock database to compute the estimators from daily prices, and it compares them with the effective spread computed by matching high-frequency trades with quotes via the TAQ database for the period 1993–2021. We find that our simulation-based results carry over to the empirical data. Our estimator is more correlated and considerably closer to the high-frequency benchmark than all other estimators in each sub-period, in each market venue, for small and large stocks, both in time-series and cross-sectional studies. Finally, we illustrate how the estimator improves inference in a variety of applications.

The second chapter studies liquidity premia where liquidity is measured by the bid-ask spread as in the seminal work by Amihud & Mendelson (1986). We find that, holding all other characteristics equal, a stock with a spread of 1.00% comes with a 0.22% larger monthly return compared to an ideal zero-spread asset. This number implies that investors expect to be compensated for transaction costs within about 4.5 months (1/0.22). The coefficient is robust across different specifications for the control variables and it remains stable across different sub-samples. When looking at seasonality, we find that the price of illiquidity spikes in January but it remains significant also in non-January months. To study how the price of illiquidity varies across firm size, we split stocks in size groups based on their market capitalization. When we use three size groups, we find that the spread is consistently priced across all groups. When we increase the number of groups to 25, we find that estimates come with larger dispersion the more we move from small-size groups to large-size groups. This suggests that liquidity is priced across all stocks, although samples restricted to larger stocks lack power to detect a significant effect. For the market portfolio (value weighted), we estimate an illiquidity premium of 0.09% per month. The equity premium is usually estimated to be around 6% in annual terms (Kroencke, 2017). Accordingly, we can attribute an economically significant share of about 20% of the historical equity premium to compensation for illiquidity. In real dollars, the illiquidity premium is 6.8 millions per stock per month. For the whole market, we estimate a premium of 4.12 billion dollars per month to compensate for illiquidity. Next, we empirically assess the predictions in Amihud & Mendelson (1986). To this end, we split stocks in groups based on the their effective bid-ask spread and we study how the price of illiquidity and the illiquidity premium vary across spread size. In line with the predictions, we find that the price of illiquidity declines as we move to higher-spread groups, while the premium increases. Finally, we formally show that the Amihud (2002) illiquidity measure can be seen as a non-linear proxy of the bid-ask spread. In summary, we find that bid-ask spreads represent an important dimension of liquidity and they have pervasive effects on asset prices.

In the last chapter, I propose a theory where prices are formed in a purely mechanical manner through trading. The theory consists of three fundamental propositions about price impact and price formation. Proposition I introduces a market clearing condition that encodes how orders are matched in modern electronic markets. This condition states that the quantity exchanged in a trade equals the integrated density of the book between zero and the peak impact. Proposition II introduces a condition of fair pricing such that both parties involved in a trade earn zero profits in the transaction. This condition implies that the peak impact relaxes to a non-zero permanent impact. Finally, Proposition III states that asset prices are the result of the accumulation of price impacts. The model produces several testable predictions on the joint relationships among trading volume, price impact, trading frequency, bid-ask spreads, order imbalance, market depth, and volatility. The theory proposed here presents a mechanical view of financial markets where trading itself is the source of price fluctuations and suggests a new research path in which the age-old question of what drives prices is replaced with what motivates people to trade in the first place.

Bibliography

Abdi, F., & Ranaldo, A. (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies*, *30*(12), 4437–4480.

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

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

Corwin, S. A., & Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance*, *67*(2), 719–760.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *Journal of Finance*, *64*(3), 1445–1477.

Holden, C. W., & Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance*, *69*(4), 1747–1785.

Jahan-Parvar, M. R., & Zikes, F. (2019). When do low-frequency measures really measure transaction costs? Working paper, SSRN 3418948.

Johann, T., & Theissen, E. (2017). The best in town: A comparative analysis of low-frequency liquidity estimators. Working paper, SSRN 2905032.

Karnaukh, N., Ranaldo, A., & Söderlind, P. (2015). Understanding FX liquidity. *Review of Financial Studies*, *28*(11), 3073–3108.

Kroencke, T. A. (2017). Asset pricing without garbage. *Journal of Finance*, *72*(1), 47–98.

Lesmond, D. A., Ogden, J. P., & Trzcinka, C. A. (1999). A new estimate of transaction costs. *Review of Financial Studies*, *12*(5), 1113–1141.

Nieto, B. (2018). Bid–ask spread estimator from high and low daily prices: Practical implementation for corporate bonds. *Journal of Empirical Finance*, *48*, 36–57.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, *39*(4), 1127–1139.

Chapter 1

Efficient Estimation of Bid-Ask Spreads from Open, High, Low, and Close Prices

The bid-ask spread is a key measure in finance, with applications ranging from asset pricing (*e.g.*, Amihud & Mendelson, 1986), to corporate finance (*e.g.*, Barclay & Smith Jr., 1988; Conroy et al., 1990), accounting research (*e.g.*, Coller & Yohn, 1997), studies on information and information asymmetry (*e.g.*, Copeland & Galai, 1983; Chung et al., 1995; Glosten & Milgrom, 1985), insider trading (*e.g.*, Chung & Charoenwong, 1998), liquidity and liquidity risk (*e.g.*, Acharya & Pedersen, 2005; Goyenko et al., 2009), and the evaluation of trading strategies and limits to arbitrage (*e.g.*, Novy-Marx & Velikov, 2016; Detzel et al., 2022), among others. However, an accurate estimate of bid-ask spreads requires high-frequency data, which are typically unavailable for international markets, asset classes other than stocks, and historical data samples. Accordingly, most studies either limit the sample to the time periods and markets of common data coverage or use proxies estimated from low-frequency data only.

Following the seminal work by Roll (1984), several approaches have been proposed to estimate bid-ask spreads from daily prices. Lesmond et al. (1999) introduce a model based on the frequency of zero returns. Hasbrouck (2009) proposes a Gibbs estimation of the Roll model. Corwin & Schultz (2012) develop a simple way to estimate bid-ask spreads from high and low prices. Abdi & Ranaldo (2017) derive an estimator that jointly considers high, low, and close prices. However, these estimators are biased when trading is infrequent and when the spread is small relative to volatility (see Corwin & Schultz, 2012; Abdi & Ranaldo, 2017; Nieto, 2018; Jahan-Parvar & Zikes, 2019; Tremacoldi-Rossi & Irwin, 2019).

In this paper, we propose an *Efficient Discrete Generalized Estimator* (EDGE) of the bid-ask spread based on price data. Our contribution to the literature is twofold.

First, we derive asymptotically unbiased bid-ask spread estimators from several combinations of Open, High, Low, and Close (OHLC) prices. As two special cases, our methodology produces the estimators in Roll (1984) and Abdi & Ranaldo (2017) with a correction term for infrequent trading. Although the previous literature has focused on continuous-time models, we show that this practice leads to a significant *downward bias* when trading becomes infrequent. Instead, our estimators remain unbiased.

Second, we combine our estimators to minimize the estimation variance and obtain an efficient estimator (EDGE). Abdi & Ranaldo (2017) point out the importance of jointly considering a wider information set of price data, rather than using close (Roll, 1984), or high-low (Corwin & Schultz, 2012) prices independently. By combining the full set of OHLC prices in an optimal way, our efficient estimator is, on average, twice as accurate as the best-performing estimators available today. The increased accuracy allows us to produce estimates closer to the (true but unobserved) bid-ask spread. Moreover, this property helps mitigate another *upward bias* that has been recently investigated by Jahan-Parvar & Zikes (2019) and Tremacoldi-Rossi & Irwin (2019). Negative estimates are usually re-set to zero to guarantee non-negativity of the transaction costs estimates in small samples (Goyenko et al., 2009; Hasbrouck, 2009; Corwin & Schultz, 2012; Karnaukh et al., 2015; Abdi & Ranaldo, 2017) and this practice leads, on average, to overstating the spread. By reducing the estimation variance, we find that EDGE naturally produces 20% fewer negative estimates compared to the alternative estimators at daily frequency.

Another advantage of EDGE is that it can be applied at any frequency. In particular, it can exploit high-frequency price data, while the other estimators struggle as trading becomes more infrequent at this time interval and their downward bias dominates the spread estimate. This property makes it possible to study bid-ask spreads in high frequency for markets that do not report bid and ask data, or when quotes are unreliable or expensive to obtain. Furthermore, EDGE admits a simple closed-form expression, does not require any matching procedure, and takes little computational time. This is particularly relevant when estimating spreads from high-frequency data, which are typically huge in size and continue to grow exponentially (Fong et al., 2017).

We compare EDGE with the seminal Roll (1984) estimator, and with Corwin & Schultz (2012) and Abdi & Ranaldo (2017), which have been shown to generally deliver the most accurate estimates in several comparative studies (Corwin & Schultz, 2012; Holden & Jacobsen, 2014; Karnaukh et al., 2015; Abdi & Ranaldo, 2017; Johann & Theissen, 2017).

In our simulation experiments, we compare the bias and variance achieved by the estimators using daily data and an estimation window ranging from one month to one year. We also run the comparison for simulations performed in high-frequency, where we use minute data and an estimation window of one hour. We find that EDGE produces the lowest bias and variance in all scenarios, indicating it is always the best choice regardless of the estimation window and the evaluation metric used by the researcher, both in low and high frequencies.

Our empirical analysis uses the CRSP U.S. stock database to compute EDGE and the other estimators from daily prices, and it compares them with the effective spread computed by matching high-frequency trades with quotes via the TAQ database for the period 1993–2021. We find that our simulation-based results carry over to the empirical data. EDGE is more correlated and considerably closer to the high-frequency benchmark than all other estimators in each sub-period, in each market venue, for small and large stocks, both in time-series and cross-sectional studies. We show that the adjustment for infrequent trading is the main driver of the improvement. Open prices provide additional information and improve the estimate of the spread further. Accordingly, when open prices are unavailable, we find that our estimator derived from close, high, and low prices is still the best option, as it outperforms all the alternative estimators that use the same information set.

To illustrate the benefits of EDGE, we provide three example applications. First, we revisit historical spread estimates, and compare the EDGE estimates from daily prices with previous approaches. We find that the other estimators overstate bid-ask spreads where they are expected to be the smallest and understate bid-ask spreads where they are expected to be the largest. The economic magnitude of this effect is large: small spreads are overstated by a factor that is several times the spread, and, depending on the estimator, large spreads are understated by a factor between 20% and 80% in relative terms. Similar, and arguably smaller, biases in effective spreads have already been shown to affect research inference (see Holden & Jacobsen, 2014; Hagströmer, 2021). Second, we study how portfolio sorts differ depending on the spread estimator used. Following Holden & Jacobsen (2014), we calculate the percentage of stock-month observations where EDGE allocates a stock to a different portfolio compared to the portfolios of each of the alternative spread estimators. We find that, on average, only 50% of the stock-months are in the same portfolios. Third, we re-examine the role of liquidity premia in explaining expected stock returns (Amihud & Mendelson, 1986; Eleswarapu & Reinganum, 1993; Amihud, 2002; Ben-Rephael et al., 2015; Amihud, 2019; Drienko et al., 2019; Harris & Amato, 2019). EDGE is significantly priced and it subsumes the explanatory power of earlier spread estimators in Fama–MacBeth regressions. We find that the monthly premium of an asset with a bid-ask spread of 1.00% is 0.22%. Put differently: Investors demand compensation for transaction costs within a holding period between four and five months. Moreover, we find that EDGE drives out the size effect of Banz (1981). This finding corroborates the idea that size is empirically priced because size is merely a proxy for liquidity (Amihud & Mendelson, 1986).

In summary, our contribution to the literature is as follows. First, we derive asymptotically unbiased bid-ask spread estimators from several combinations of open, high, low, and close prices. Second, we provide the optimal way of combining the estimators to minimize the estimation variance and obtain an efficient estimator (EDGE). Finally, we show that our efficient estimator has important implications in applied research. To guarantee the reproducibility of our work, we make available software for the R statistical environment (R Core Team, 2020) that implements all the results in this paper. EDGE is available in a variety of programming languages.[1]

1.1 Methodology

For a given asset, the percent effective spread S on a trade is defined as:

$$S = \frac{2D(P - \tilde{P})}{\tilde{P}}\,, \tag{1.1}$$

where P is the (observed) transaction price, $\tilde{P}$ is the (unobserved) fundamental value, and D is a direction of trade indicator taking the value $+1$ for buyer-initiated trades, and -1 for seller-initiated trades. In logarithmic prices, Equation (1.1) becomes:

$$p = \tilde{p} + Z\,, \tag{1.2}$$

where $Z = S(B - 0.5)$ and B takes the value 1 for buyer-initiated trades, and 0 for seller-initiated trades. Here we make two assumptions:

1. The (log-)returns of the fundamental value $\tilde{p}$ are uncorrelated.

[1]The code implementing the estimator is available at `https://github.com/eguidotti/bidask`.

2. Buys and sells at the open and close are equally likely.

Assumption 1 states that the fundamental value of a security fluctuates randomly, on the lines of Roll (1984). This assumption of zero autocorrelation is less restrictive than independence, which is popular among previous contributions. In particular, we do not rule out the possibility that the squared returns are autocorrelated (*i.e.*, time-varying volatility and volatility clustering). Unlike other low-frequency estimators, we do not need to assume any specific model for the fundamental value $\tilde{p}$ (*e.g.*, Corwin & Schultz, 2012; Abdi & Ranaldo, 2017). Unlike other high-frequency estimators, we do not need to assume that the fundamental value $\tilde{p}$ is represented by the quoted midpoint (*e.g.*, Holden & Jacobsen, 2014).

Assumption 2 states that B is a Bernoulli random variable with probability of success 50%, and our framework is expected to be robust to this assumption (see Section 1.1.2).

Finally, the spread can be time-varying and random within a period (*e.g.*, day) or between different periods. In this case, Appendix 1.A.1 shows that all the estimators that we derive are formally estimators for the root-mean-square spread.

1.1.1 Derivation of Discrete Generalized Estimators

We derive our bid-ask spread estimators by computing the covariance between subsequent price changes from several combinations of open, high, low, and close prices. For simplicity, we illustrate the derivation of an estimator based on close prices only. The other cases involve more tedious calculations, but they follow the same logic. All the calculations are reported in Appendix 1.A.1.

Let us define c_t as the log-price at the end of a trading time interval (*e.g.*, closing of the day). We compute the covariance between the log-return $c_t - c_{t-1}$ to its first lag by replacing the observed log-prices c_t with the actual (but unobserved) log-prices $\tilde{c}_t$ in Equation (1.2):

$$\begin{aligned} \mathbb{C}\mathrm{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] &= \mathbb{C}\mathrm{ov}[\tilde{c}_t - \tilde{c}_{t-1}, \tilde{c}_{t-1} - \tilde{c}_{t-2}] \\ &+ \mathbb{C}\mathrm{ov}[\tilde{c}_t - \tilde{c}_{t-1}, Z_{t-1} - Z_{t-2}] \\ &+ \mathbb{C}\mathrm{ov}[Z_t - Z_{t-1}, \tilde{c}_{t-1} - \tilde{c}_{t-2}] \\ &+ \mathbb{C}\mathrm{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}] \\ &= \mathbb{C}\mathrm{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}]\,, \end{aligned} \tag{1.3}$$

where the first three terms are zero because the actual returns are not autocorrelated by assumption and they are independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned} \mathbb{C}\mathrm{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}] &= \mathbb{C}\mathrm{ov}[Z_t, Z_{t-1}] \\ &+ \mathbb{C}\mathrm{ov}[Z_t, -Z_{t-2}] \\ &+ \mathbb{C}\mathrm{ov}[-Z_{t-1}, Z_{t-1}] \\ &+ \mathbb{C}\mathrm{ov}[-Z_{t-1}, -Z_{t-2}]\,. \end{aligned} \tag{1.4}$$

So far, the literature has assumed that the covariance between bid-ask bounces at different times is zero, so that Equation (1.4) reduces to $\mathbb{C}\mathrm{ov}[-Z_{t-1}, Z_{t-1}] = -\mathbb{V}[Z]$. The reason is that, if there is always at least one trade between time t and time s (*e.g.*, continuous-time models), then the random variable Z_t is independent from Z_s for any $s \neq t$. However, we point out that, in practice, this is not true. For instance, one trade originates two subsequent closing prices if no trade has occurred in the second period or many correlated transactions are executed at the same price. In these circumstances, bid-ask bounces at different times are not independent, as they are originated by the same trade. Assuming that the only non-vanishing term in Equation (1.4) is $-\mathbb{V}[Z]$ will lead to a biased estimator of the bid-ask spread. We add to the literature by deriving a generalized formula that computes all the terms explicitly and, thus, it allows different prices to incorporate the same bid-ask bounce, as it is often the case of periods with few trades or no trades at all.[2] By the covariance decomposition formula (see Appendix 1.A.1) we have:

$$\mathbb{C}\text{ov}[Z_t, Z_{t-1}] = \mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}] \,, \tag{1.5}$$

where $\mathbb{P}[Z_t = Z_{t-1}]$ is the probability that Z_t and Z_{t-1} are the same bid-ask bounce. The same holds for:

$$\mathbb{C}\text{ov}[-Z_{t-1}, -Z_{t-2}] = \mathbb{V}[Z]\mathbb{P}[Z_{t-1} = Z_{t-2}] \,, \tag{1.6}$$

and, over two time periods, we have:

$$\mathbb{C}\text{ov}[Z_t, -Z_{t-2}] = -\mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}]\mathbb{P}[Z_{t-1} = Z_{t-2}] \,. \tag{1.7}$$

We estimate the probability that two subsequent closing prices incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{c=c}$) in which the two prices are equal:[3]

$$\nu_{c=c} \mathrel{\hat{=}} \mathbb{E}[c_t = c_{t-1}] = \mathbb{P}[Z_t = Z_{t-1}] = \mathbb{P}[Z_{t-1} = Z_{t-2}] \,. \tag{1.8}$$

We can now compute the covariances in Equations (1.3)–(1.4) by substituting the terms in Equations (1.5)–(1.8):

$$\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = -\mathbb{V}[Z](1 - 2\nu_{c=c} + \nu_{c=c}^2) = -\mathbb{V}[Z](1 - \nu_{c=c})^2 \,. \tag{1.9}$$

The final formula is obtained by expressing $\mathbb{V}[Z]$ in function of the squared spread S^2 (see Appendix 1.A.1):

$$\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = -\frac{S^2}{4}(1 - \nu_{c=c})^2. \tag{1.10}$$

[2]For instance: a) the market reports the previous closing price when there is no trade; b) one trade can originate contemporaneously the open, high, low, and close prices if it is the only trade in the period; c) one trade can originate both the high (low) and close (open) price if the closing (opening) trade is selected as the highest (lowest) price. In these cases, different prices on the market are originated by the same trade and, thus, they incorporate the same bid-ask bounce.

[3]This assumes that only close prices are available and it is for illustration. In the general case, we refine the estimate by using high and low prices. We estimate the probability that two subsequent closing prices incorporate the same bid-ask bounce by counting the fraction of times in which both the high and the low prices at time t are equal to the closing price at time $t-1$ (see Appendix 1.A.1).

Equation (1.10) can be easily solved for S^2:

$$S^2 = -\frac{4\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}]}{(1-\nu_{c=c})^2}\,, \tag{1.11}$$

which is an unbiased estimator of the bid-ask spread based on closing prices.

Exploiting additional information from open, high and low prices is expected to provide a more efficient estimator (see Abdi & Ranaldo, 2017). Following this reasoning, we derive unbiased bid-ask spread estimators using open, high, low, and close prices, as well as combinations of these prices. The calculations are provided in Appendix 1.A.1.

In Table 1.1, we summarize the various estimators. Two special cases are obtained. When $\nu_{c=c} = 0$ (*i.e.*, subsequent closing prices are always different), the C estimator is identical to the bid-ask estimator of Roll (1984). Similarly, when $\nu_{c=h,l} = \nu_{h=l=c} = 0$ (*i.e.*, the closing price is never selected as the highest or lowest price), the CHL estimator is identical to the bid-ask estimator of Abdi & Ranaldo (2017).[4]

The first innovation of the estimators displayed in Table 1.1 is that they account for infrequent trading. While the estimators in Roll (1984), Corwin & Schultz (2012), and Abdi & Ranaldo (2017) lead to biased results when trading is infrequent, our estimators remain unbiased. Therefore, they can be applied to a wide range of asset classes, in liquid and illiquid market segments, at low or high frequencies.

The second innovation of the estimators shown in Table 1.1 is that they extend over the full set of information by jointly considering open, high, low, and close prices. The remaining open question is which of these estimators, or a combination of estimators, should be chosen to obtain the best possible (efficient) estimator.

[Insert Table 1.1 about here.]

1.1.2 The Efficient Discrete Generalized Estimator (EDGE)

In this section, we combine our generalized estimators to minimize the estimation variance and obtain our efficient estimator (EDGE). To this end, we follow three steps. First, we show that each estimator in Table 1.1 can be written as a moment condition so that the asymptotically efficient estimator is obtained by applying the Generalized Methods of Moments (GMM) (Hansen, 1982) (Appendix 1.A.2). Second, we include prior knowledge to improve the efficiency in small samples (Appendix 1.A.2). Third, we provide an estimator for k in Table 1.1, which accounts for high (low) prices to be usually buyer (seller) initiated (Appendix 1.A.2).

Our *Efficient Discrete Generalized Estimator* (EDGE) of the bid-ask spread is given by:

$$S^2 = -4\frac{w_1\mathbb{E}[X_1] + w_2\mathbb{E}[X_2]}{(1-k\nu_{o=h,l}) + (1-\nu_{h=l=c})(1-k\nu_{c=h,l})}\,, \tag{1.12}$$

where $\nu_{o=h,l}$, $\nu_{c=h,l}$, and $\nu_{h=l=c}$ are given in Table 1.1; k is set to $k = 4w_1w_2$; and the vectors X_1, X_2, together with the weights w_1, w_2, are provided below:

$$\begin{aligned} X_{1,t} &= (\eta_t - o_t)(o_t - \eta_{t-1}) + (\eta_t - c_{t-1})(c_{t-1} - \eta_{t-1})\,, \\ X_{2,t} &= (\eta_t - o_t)(o_t - c_{t-1}) + (o_t - c_{t-1})(c_{t-1} - \eta_{t-1})\,, \end{aligned} \tag{1.13}$$

[4] This can be easily seen by computing $\mathbb{C}\text{ov}[r_t, r_{t-1}] = \mathbb{E}[r_t r_{t-1}]$ for zero-mean log-returns.

$$w_1 = \frac{\mathbb{V}[X_2]}{\mathbb{V}[X_1] + \mathbb{V}[X_2]}\,, \quad w_2 = \frac{\mathbb{V}[X_1]}{\mathbb{V}[X_1] + \mathbb{V}[X_2]} \quad . \tag{1.14}$$

For estimation, the usual sample counterparts replace the expectations and variances, respectively. We expect EDGE to provide superior performance compared to the estimators by Roll (1984), Corwin & Schultz (2012), and Abdi & Ranaldo (2017) as it is derived under more general conditions and it takes advantage of the whole information set by jointly considering the full set of opening, high, low, and closing price data in an optimal way.

Negative Estimates

The estimate $\hat{S}^2$ in Equation (1.12) may become negative in small samples. To guarantee non-negativity of the transaction costs estimate, we follow the common approach of truncating negative values (Goyenko et al., 2009; Hasbrouck, 2009; Karnaukh et al., 2015):

$$\hat{S} = \sqrt{\max\left\{0\,,\hat{S}^2\right\}}\,. \tag{1.15}$$

This practice introduces an upward bias in small samples because negative estimates are set to zero (Jahan-Parvar & Zikes, 2019; Tremacoldi-Rossi & Irwin, 2019). By minimizing the estimation variance, we expect EDGE to naturally produce fewer negative estimates and minimize this small-sample bias.

Robustness

When considering that buyer-initiated trades may occur more frequently (infrequently) than seller-initiated trades, the variance of Z becomes $S^2b(1-b)$ instead of $S^2/4$ as shown in Appendix 1.A.1. By using $S^2b(1-b)$ instead of $S^2/4$ in Appendix 1.A.1, it can be seen that all the equations in Table 1.1 hold more generally by substituting the coefficient 4 with $1/[b(1-b)]$ in the right-hand side of the equations, where b is the probability to observe a buyer-initiated trade. For instance, by choosing $b = 0.6$, we obtain a coefficient for S^2 of 4.17, instead of 4, that translates in a coefficient for the spread S equal to $\sqrt{4.17} \approx 2.04$, very close to $\sqrt{4} = 2$. In other words, as long as $0 \ll b \ll 1$, all the formulas in Table 1.1, and in particular Equation (1.12) hold, at least approximately, even if buys and sells are not equally likely.

All the estimators further allow for high-frequency serial dependence in the trade directions (Hasbrouck & Ho, 1987), as any autocorrelation $\delta < 1$ between subsequent trade directions vanishes exponentially fast (δ^n) as the lags (n) increase, producing a negligible serial dependence in the trade directions of the open and close prices.

1.2 Simulation Study

In this section, we perform a Monte Carlo study to assess the accuracy and robustness of the EDGE in Equations (1.12)–(1.15). We compare the results with the seminal Roll (1984) estimator, and with the estimators proposed more recently by Corwin & Schultz (2012) and Abdi & Ranaldo (2017). With ROLL, we refer to the Roll (1984) estimator. The other two papers define at least two versions of their estimators that handle negative spread estimates in different ways. The first version simply sets negative estimates to zero and it offers the most natural benchmark for our estimator. We refer to these versions as the CS (Corwin & Schultz, 2012) and AR (Abdi & Ranaldo, 2017) estimators, respectively.[5] The second version estimates spreads separately for each pair of consecutive periods, sets them to zero when necessary, and calculates the final estimate as the average across all the two-period estimates. We refer to these versions as the CS2 (Corwin & Schultz, 2012) and AR2 (Abdi & Ranaldo, 2017) estimators, respectively.[6] The CS and CS2 estimators are adjusted for overnight returns as described in Corwin & Schultz (2012).

1.2.1 Setup

For ease of comparison, we use the simulation setup of Corwin & Schultz (2012) that is also used in Abdi & Ranaldo (2017).

Low Frequency

We simulate 10,000 months of stock data where each month consists of 21 days and each day consists of 390 minutes. For each minute of the day, the true value of the stock price, P_m, is simulated as $P_m = P_{m-1}e^{\sigma x}$, where σ is the standard deviation per minute and x is a random draw from a standard Gaussian distribution. The daily standard deviation equals 3%, and the standard deviation per minute equals 3% divided by $\sqrt{390}$. In each simulation, stock prices are assumed to be observed each minute with a given probability. The bid (ask) for each minute is defined as P_m multiplied by one minus (plus) half the assumed bid-ask spread, and we assume a 50% chance that a bid (ask) is observed. Daily high and low prices equal the highest and lowest prices observed during the day. Open and close prices equal the first and the last price observed in the day. If no trade is observed for a given day, then the closing price of the previous day is used as the open, high, low, and close prices for that day.

High Frequency

Similar to the setup above, we run high-frequency simulations, in this case consisting of 252 8-hour stock-days, and where each day consists of $8 \times 60 \times 60 = 28{,}800$ seconds. The standard deviation per second equals 3% divided by $\sqrt{28{,}800}$. Stock prices are assumed to be observed each second with a given probability, and with a 50% chance that a bid (ask) is observed. The high and low prices per minute equal the highest and lowest prices observed during the minute. Open and close prices equal the first and the last price observed in the minute. If no trade is observed for a given minute, then the closing price of the previous minute is used as the open, high, low, and close prices for that minute.

[5] This is the *Monthly* version used in the original papers.
[6] This is the *2-Day* version used in the original papers.

1.2.2 Results

In Table 1.2, we report monthly estimates obtained by the daily prices simulated in Section 1.2.1. Panel A shows the simulation results under near-ideal conditions, where there are 390 trades per day and overnight returns are not included. In these simulations, EDGE is twice more precise than AR and CS, and four times more precise than the ROLL estimator. For example, for a true spread of 0.50%, the EDGE estimate is 0.44% with a standard deviation of 0.33%, while the AR (CS) estimate is 0.70% (0.60%) with a standard deviation of 0.77% (0.49%). By estimating spreads as large as 1.21%, 1.44%, 1.41%, respectively, AR2, CS2, and ROLL are not accurate for small spreads as already observed by Tremacoldi-Rossi & Irwin (2019), and also in the original papers. For larger spreads, the estimators become more similar but with EDGE always achieving the most precise estimates. In Panel B, we introduce infrequent trading and overnight jumps in the simulations. These results highlight the robustness of EDGE compared to the other estimators. We find that EDGE outperforms the CS and the ROLL estimators with better accuracy and lower bias in all scenarios. EDGE performs similarly to the AR estimator for the smallest considered spread (0.50%), and it is more accurate and precise in all other scenarios.

[Insert Table 1.2 about here.]

In the Appendix (Table 1.9), we replicate the experiment by Jahan-Parvar & Zikes (2019), who compare the bias of the estimators for tiny spreads when one year of daily data is used. We find that EDGE is the only estimator able to consistently estimate spreads as small as 0.10% under near-ideal conditions, while CS, AR, and ROLL produce upward-biased estimates of 0.20%, 0.34%, and 0.66%, respectively. The results deteriorate when overnight returns are included in the simulations, but EDGE always ameliorates the upward bias and produces consistent estimates for spreads equal to 0.30% and larger.

In Table 1.8, we extend the comparison to the high-frequency setting described in Section 1.2.1. Under near-ideal conditions, we find that all the estimators perform similarly, but with EDGE always achieving the best accuracy. In the infrequent trading setting, the performance gap between EDGE and the other estimators widens significantly. EDGE is the only reliable estimator for the simulation experiment that mimics intraday data.

The remainder of this section is dedicated to a deeper comparison of the estimators from several perspectives.

Bias

In Figure 1.1, we study the bias of the estimators as a function of the average number of trades per day. We simulate the low-frequency setting in Section 1.2.1 where the probability of observing a trade ranges from 0.5% to 100% so that the corresponding expected number of trades per day ranges from 2 to 390. We use a constant spread of 1% and compare the results obtained with EDGE, AR, and CS. CS is significantly biased and converges slowly to the true spread as the expected number of trades per day increases. AR converges faster, but it is still biased when the expected number of daily trades is below 30. EDGE produces unbiased estimates regardless of the number of trades, suggesting it works well in practice even in the case of illiquid assets or in high frequency when only few trades are observed per period. The results for CS2 and AR2 are not reported as they are significantly biased, even for a very large number of trades per day, as shown in Table 1.2 and already documented in the original papers. In the Appendix (Figure 1.12), we extend the comparison to the high-frequency setting described in Section 1.2.1, from which the same conclusions can be drawn.

[Insert Figure 1.1 about here.]

Variance

In Figure 1.2, we study the standard deviation of the bid-ask spread estimators depending on the magnitude of the spread. To this end, we run the simulations described in Section 1.2.1, estimate the spread for each month, and compute the standard deviation of the estimates. The procedure is repeated for several levels of the spread. These simulations use 390 trades per day so that all the estimators are unbiased (see Figure 1.1) and the minimum-variance estimator coincides with the best estimator in the usual root mean squared error sense. We notice that CS is preferable to AR for small spreads, while AR achieves better performance for larger spreads. In both cases, EDGE provides the most precise estimates with a standard deviation lower than the other approaches across low and large spreads. In the Appendix (Figure 1.13), we extend the comparison to the high-frequency setting described in Section 1.2.1, from which the same conclusions can be drawn.

[Insert Figure 1.2 about here.]

Negative Estimates

Jahan-Parvar & Zikes (2019) and Tremacoldi-Rossi & Irwin (2019) document that the practice of resetting negative estimates to zero leads, on average, to overstating the spread. In this section, we show that EDGE naturally produces fewer negative estimates with respect to all other estimators, and the fraction of negative estimates can be further reduced by increasing the estimation window.

In Figure 1.3, we study how the proportion of zero estimates varies in function of the sample size. To this end, we run the simulations described in Section 1.2.1, estimate the spread using an estimation window ranging from one month to one year, and compute the corresponding percentage of zero estimates that we obtain. The simulations use a constant spread of 1%, a 10% probability of observing a trade (for an average of 39 trades per day), and an overnight return normally distributed with mean zero and standard deviation equal to half of the daily volatility. We notice how the ROLL estimator produces a large number of zero estimates (about 40% of the time) even when using one year of daily data to compute the spread. AR and CS exhibit similar behavior, producing non-positive estimates between 20% and 30% of the time with a one-year estimation window. Instead, EDGE significantly reduces the frequency of zero estimates as the sample size increases, reaching a fraction lower than 5% for a one-year estimation window.

[Insert Figure 1.3 about here.]

1.2.3 Stress Test

In this section, we report the simulation results in which we include different imperfections simultaneously, such as adding overnight jumps that proxy for non-trading periods, allowing the probability to observe a trade to vary over time, and assuming a time-varying random spread.

In Figure 1.4, we simulate this scenario for 10,000 stock-months with 21 days in each month. For each day, we use the previous year to estimate the spread. The estimates are benchmarked with the true spread and the average number of trades in the previous year. One clear result emerges. EDGE exhibits the smallest variance, and it is also able to disentangle the spread dynamics from the expected number of trades per day. AR, AR2, CS, and CS2 considerably underestimate the spread in periods when the number of trades per day is low. The researcher should be careful when applying these estimators in practice, as changes in trading frequency are likely to be artificially reflected on changes in estimated spreads. For example, spread estimates in the 1930s will not be directly comparable with estimates following the introduction of electronic trading that significantly increased the trading frequency. EDGE allows for a consistent comparison regardless of changes in the trading dynamics. Finally, we note that AR2 and CS2 exhibit a lower variance but are more biased with respect to AR and CS. The ROLL estimator is affected by a large variance that makes practical estimation hard in practice.

[Insert Figure 1.4 about here.]

1.3 Empirical Results

In this section, we investigate the performance of the estimators on empirical data. To evaluate the performance, we first need to define the ground truth, that is, the spread that serves as the benchmark for the evaluation. Following the literature, we use the effective spread obtained by matching high-frequency trade and quote data to evaluate the performance of the various estimators that only require commonly available daily price data.

1.3.1 Data

To compute bid-ask spread estimates (*i.e.*, EDGE, AR, AR2, CS, CS2, ROLL), we obtain daily prices from the CRSP US Stock Database in the period 1926–2021. To ensure that all the estimates are obtained from transaction prices only, we keep the observations for which the open, high, low, and close prices are directly available.[7] We select all NYSE, AMEX, and NASDAQ stocks with CRSP share codes of 10 or 11 (*i.e.*, U.S. common

[7]If transaction prices are not available, CRSP reports quotes derived from bid and ask prices. These values are marked in CRSP by a dash in front of the price. We consider these non-transaction-based prices as missing values. Then, we drop the days where the high, low, or close price is missing. Finally, we drop a few observations where the open or close prices are outside the high-low range or where the low price is higher than the high price. Open prices are missing in CRSP from July 1962 through June 1992.

shares). No other data pre-processing is performed to maintain all the complexity of empirical data and the infrequently traded stocks with only a few monthly observations.

To compute the benchmark effective spread, we rely on the Trades and Quotes (TAQ) database available in 1993–2021.[8] The benchmark spreads are the average effective spreads during the day and are obtained via the Wharton Research Data Services (WRDS) using Monthly TAQ for 1993–2003 and Daily TAQ from 2004 onward, according to the methodology described in Holden & Jacobsen (2014).[9] We refer to this measure as HJ (Holden & Jacobsen, 2014). We match CRSP and TAQ data using CUSIP identifiers and tickers.[10] Our identification strategy allows us to match about 99% of the stocks in CRSP.

For each month, we estimate the spread with EDGE, AR, AR2, CS, CS2, and ROLL and drop the monthly estimate for all the estimators if it is missing for any of them. We also include CHL (Table 1.1) in the comparison to illustrate the incremental contribution from using open prices versus the infrequent trading adjustment in EDGE. The monthly benchmark is computed as the average of the HJ effective spread within the month. Our CRSP-TAQ merged sample consists of about 1.6 million stock-month spread estimates for each estimator in the period 1993–2021.

Table 1.3 displays the correlation among the estimators and Table 1.4 provides summary statistics. We report the mean, median, and standard deviation for the estimates and the effective spread benchmark. We further report the fraction of non-positive estimates and several evaluation metrics for the positive estimates. We notice how EDGE achieves the smallest fraction of zero estimates (25.06%), the highest correlation with the benchmark (81.82%), and the lowest MAPE and RMSE.[11] The comparison between CHL and AR shows that adjusting for infrequent trading is essential to improve the correlation with the benchmark. The comparison between CHL and EDGE shows that open prices do provide additional information to further improve the spread estimates. The remainder of this section is dedicated to a deeper comparison across the estimators in a cross-sectional, time-series, and panel-data setting.[12]

[8]Our sample starts from May 1993 as we find that TAQ is affected by a strong selection bias before that date, when it lists only the most liquid stocks.

[9]We use the effective spreads computed via the methodology in Holden & Jacobsen (2014) that are available to download from the WRDS Intraday Indicators. We also compute the benchmark effective spread using the weighted-midpoint as described in Hagströmer (2021). First, we calculate effective spreads using the quoted midpoint as in Holden & Jacobsen (2014) for all stock-days in the Daily TAQ database in the period 2004–2021: Our estimates achieve a correlation of 99.5% with the estimates pre-computed by WRDS. Then, we replace the midpoint with the weighted-midpoint to generate the effective spreads described in Hagströmer (2021). We find that the correlation between the midpoint and the weight-midpoint effective spreads is above 98%. We have evaluated the estimators using both benchmarks and the results are fully consistent. We use the benchmark estimates by Holden & Jacobsen (2014) as they are more easily available also for the Monthly TAQ database in the period 1993–2003, where the NBBO is not directly available.

[10]We reconstruct the time series of CUSIPs for each KYPERMNO in CRSP. Similarly, we reconstruct the time series of TICKERs for each KYPERMNO in CRSP. Then, we compute the time series of CUSIPs for each SYMBOL in TAQ using the Monthy TAQ Master files for 1993–2009 and the Daily TAQ Master files for 2010–2021. Finally, we merge the (daily) datasets by matching observations with the same date, with the same CUSIP, and where the TAQ's SYMBOL is equal to the TICKER in CRSP.

[11]We recall that AR2 and CS2 tend to avoid zero estimates by construction. The MAPE and RMSE are computed on the log-spreads as described in Appendix 1.A.4.

[12]AR2 and CS2 perform worse than AR and CS and we do not report them in the following analyses.

[Insert Tables 1.3 and 1.4 about here.]

1.3.2 Cross-Sectional Correlation

Looking at cross-sectional correlations on a month-by-month basis allows us to evaluate the estimators' ability to capture the cross-sectional distribution of spreads in different time periods. Given the effective spread benchmark $S_{i,t}$ for stock i at time t and the corresponding estimate $\hat{S}_{i,t}$, we compute the cross-sectional correlation at time t as $\rho_t = \mathbb{C}\mathrm{or}_i[S_{i,t}, \hat{S}_{i,t}]$ using all $\hat{S}_{i,t}$ that are positive. The month-by-month cross-sectional correlations for the various estimators are displayed in Figure 1.5. We see that the correlation between EDGE and the effective spread benchmark is consistently higher than the correlations achieved by any other estimator throughout the whole period considered in the analysis.

[Insert Figure 1.5 about here.]

1.3.3 Time-Series Correlation

Looking at time-series correlations on a stock-by-stock basis allows us to evaluate the ability of the estimators to capture the time-series distribution of spreads for different kinds of stocks. To this end, we split all stocks in deciles based on their market capitalization.[13] Then, given the effective spread benchmark $S_{i,t}$ for stock i at time t and the corresponding estimate $\hat{S}_{i,t}$, we compute the time-series correlation for decile d as $\rho_d = \mathbb{C}\mathrm{or}_{t,i\in d}[S_{i,t}, \hat{S}_{i,t}]$ using all $\hat{S}_{i,t}$ that are positive. The time-series correlations for each decile obtained with the various estimators are displayed in Figure 1.6. We see that the correlation between EDGE and the effective spread benchmark is consistently higher than the correlations achieved by any other estimator for all types of stocks.

[Insert Figure 1.6 about here.]

1.3.4 Panel-Data Correlation

Next, we analyze the performances across four dimensions: market venues, time periods, market capitalization, and spread size. When analyzing market venues, the groups correspond to NYSE, AMEX, and NASDAQ. For the time periods, we use those defined in Corwin & Schultz (2012) and Abdi & Ranaldo (2017). In addition, we extend the sample and include the more recent sub-period 2016–2021. For market capitalizations, we split the stocks in quintiles using the same procedure described in Section 1.3.2. For spread sizes, we split the stocks in quintiles based on the average effective spread throughout the life of the stock. Then, given the effective spread benchmark $S_{i,t}$ for stock i at time t and the corresponding estimate $\hat{S}_{i,t}$, we compute the correlation for group g as $\rho_g = \mathbb{C}\mathrm{or}_{(i,t)\in g}[S_{i,t}, \hat{S}_{i,t}]$ using all $\hat{S}_{i,t}$ that are positive.

The results are summarized in Table 1.5 for market venues (Panel A), time periods (Panel B), market capitalization (Panel C), and spread size (Panel D). One clear result

[13] The size deciles are sorted by increasing market capitalization of each stock as its last listing date on CRSP, as defined in Corwin & Schultz (2012) and Abdi & Ranaldo (2017).

emerges: EDGE outperforms all the alternative estimators in each market venue, subperiod, market capitalization, and spread size by consistently achieving the highest correlation with the effective spread benchmark and the lowest fraction of zero estimates.[14] The comparison between EDGE and CHL shows that open prices do provide additional information to further improve the spread estimates. When open prices are not available, our CHL estimator becomes the best option, as it outperforms all the alternative estimators that use the same information set.

[Insert Table 1.5 about here.]

1.4 Applications

To demonstrate the potential benefits of EDGE, we provide three representative examples. The first revisits historical spread estimates since 1926. The second studies how portfolio sorts differ depending on the spread estimator used. The third illustrates how EDGE can add to the asset pricing literature and the study of liquidity premia.

1.4.1 Historical Spread Estimates

Using CRSP data since 1926, we construct, for each month, three portfolios based on size according to the following procedure. First, we sort the stocks based on their market capitalization at the end of each month. Then, we select small-caps, mid-caps, and large-caps using the common 50th and 80th percentiles as breakpoints. Finally, we estimate monthly spreads from daily prices using several methods and track the average spread of the three portfolios in 1926–1992 (CRSP sample) and 1993–2021 (CRSP-TAQ merged sample).[15]

[Insert Figure 1.7 about here.]

The results are reported in Figure 1.7. As expected, the results show that smaller stocks tend to have higher spreads than larger stocks. From the recent sample (CRSP-TAQ), we conclude that EDGE closely follows the effective spread benchmark. CS and AR tend to underestimate the spread, particularly for small-cap stocks, mirroring the fact that these estimators are biased when trading is infrequent. In the arguably less liquid historical sample prior to 1993, we find that the gap between EDGE and the alternative estimators further widens. EDGE is by a factor of two larger than AR, and the difference is even more pronounced compared to CS. Given our benchmark result from the recent sample, we conjecture that the alternative estimators considerably underestimate the effective spread in the early sample.

[14] In Appendix 1.A.4, we also provide the comparison on MAPE and RMSE and extend the estimation window to one year. EDGE consistently achieves the highest correlation, lowest MAPE and RMSE for each sample size and evaluation metric. EDGE also achieves the best results in the estimation of first differences instead of spread levels (not reported).

[15] When EDGE cannot be computed, we use the CHL estimator in Table 1.1 that does not need open prices. We recall that open prices are missing in CRSP from July 1962 through June 1992.

[Insert Figure 1.8 about here.]

In Figure 1.8, we report the bias as a function of the spread. In the sample after 1993, the bias is computed as the average difference between the spread estimate and the HJ benchmark (left panel). Here we find that the ROLL estimator overstates small spreads by up to 100bps (reset-to-zero bias) while, at the same time, it understates large spreads (infrequent-trading bias). AR (AR2) and CS (CS2) decrease the overestimation of small spreads at the cost of increasing the underestimation of large spreads. For instance, when the benchmark spread is 5%, AR exhibits a downward bias of 100bps and CS of 350bps. In the sample period 1926–1993 (when TAQ data are not available), we use EDGE (instead of HJ) as the benchmark to compute the bias (right panel). These results confirm the behavior observed in the recent sample: Other estimators overstate bid-ask spreads where they are expected to be the smallest and understate bid-ask spreads where they are expected to be the largest. The economic magnitude of this effect is large: small spreads are overstated by a factor that is several times the spread, and large spreads are understated by a factor between 20%–40% (AR) and up to 60%–80% (CS) in relative terms. Similar—and arguably smaller—biases in effective spreads have already been shown to affect research inference (see Holden & Jacobsen, 2014; Hagströmer, 2021).

Instead, EDGE exhibits a constant upward bias of around 25bps (Figure 1.8, left panel). However, this type of bias is considerably less problematic. First, in regression analyses, a constant bias in the regressor is absorbed by the intercept and does not bias the regression coefficient.[16] Second, the bias of 25bps is less relevant for larger spreads, where it translates into a smaller overestimation of transaction costs in relative terms. Third, the EDGE bias is a small-sample bias (see Section 1.1.2) and can be arbitrarily decreased by increasing the number of observations for the estimation. When the bid-ask spread is expected to be tiny and highly accurate estimates are needed, high-frequency prices may be used to increase the sample size and improve the estimation accuracy.[17] In this case, we stress that the number of trades observed per time interval shrinks proportionally. As a result, it becomes increasingly important to apply an estimator that is not biased when trading becomes more and more infrequent.

To illustrate how EDGE can estimate unbiased tiny spreads using intraday prices, we proceed as follows. First, we aggregate high-frequency trades into open, high, low, and close prices every minute. Then, we estimate the spread with EDGE from the minute data for each day. This procedure is applied to the Daily TAQ database, consisting of about one hundred billion trades. As EDGE does not involve any compute-intensive operation, we are able to process millions of trades per second and to produce the spread estimates in about one day. Our sample consists of 30.4 million stock-day spread estimates. The monthly estimates are computed as the average of the daily estimates within the month.[18]

[16] Let $\hat{S}_i = S_i + S_0$ be the spread estimate for stock i, where S_i is the true value of the spread and S_0 is a constant bias. Then, the regression $y_i = \beta S_i + \alpha$ is equivalent to $y = \beta(\hat{S}_i - S_0) + \alpha = \beta \hat{S}_i + \alpha'$ with $\alpha' = \alpha - \beta S_0$.

[17] While the variance component of an asset return is proportional to the return interval, the spread component is not (Corwin & Schultz, 2012). Hence, we can rely on high-frequency prices to reduce the asset variance without altering the spread component and achieve a better signal-to-noise ratio to improve the spread estimate.

[18] We use Daily TAQ as recommended by Holden & Jacobsen (2014). However, we find that the

Figure 1.9 zooms into the large-caps stocks of Figure 1.8 and it compares the monthly estimates obtained with (a) the HJ benchmark spread, (b) EDGE from daily prices, and (c) EDGE from minute prices. While the estimates from daily prices are upward biased, the estimates from minute prices estimate tiny spreads with a high degree of accuracy and the upward bias shrinks to zero. For instance, the average spread is 4.9bps (0.049%) according to HJ, and it is 5.2bps (0.052%) when estimated with EDGE from minute prices.

[Insert Figure 1.9 about here.]

1.4.2 Ranking Differences

We follow Holden & Jacobsen (2014) and investigate how portfolio sorts are affected by choice of the spread estimator. Portfolio sorts are common in a wide range of applications in empirical finance and accounting. One reason is that sorts are robust to a bias in the level of the spread estimate. However, sorts are still affected by the estimation variance. Accordingly, it can be expected that EDGE provides more accurate rankings compared to other estimators from daily data.

To quantify the difference between rankings, we sort stocks into quintiles based on their spread value each month. Quintile 1 contains the smallest spreads, and quintile 5 contains the largest spreads in the month. Following the methodology of Holden & Jacobsen (2014), we then compute the percentage of stock-months where EDGE allocates a stock to a different quintile compared to the alternative spread estimators. For example, if a stock is ranked 2 by ROLL, but has a rank of 1 according to EDGE, the rank difference is $+1$, for a given stock-month. Because there are five quintiles, the rank difference ranges from -4 to $+4$.

Table 1.6 reports the percentage of stock-months where AR (AR2), CS (CS2), and ROLL allocate a stock to a lower or higher quintile compared to the quintile portfolio obtained with EDGE. Using AR, we find that only 58.2% of the stock-months are in the same portfolios of EDGE. The agreement further decreases for AR2 (48.5%), ROLL (44.6%), and CS and CS2 (33.9%). On average, more than 50% of stock-months disagree, with a percentage ranging from 20.6% (AR) to 29.1% (CS) placed in lower quintiles, and from 21.2% (AR) to 37.7% (CS2) placed in higher quintiles. Even though they are rare, it is notable that rank differences of the maximum four steps exist. Such cases indicate that a stock that is ranked highest (lowest) according to one spread estimator is ranked lowest (highest) according to EDGE.

The magnitude of ranking differences that we document is well comparable to the results reported in the study by Holden & Jacobsen (2014), which compares different high-frequency estimators based on transaction and quote data. We therefore conclude that different low-frequency estimators have a comparably large impact on a wide range of research areas.

---

correlation between the EDGE spread estimates obtained with Monthly or Daily TAQ in the overlapping period (2003–2014) is 98%. This suggests that EDGE is not affected by second (Monthly TAQ) or millisecond (Daily TAQ) timestamps, and it is somewhat more robust than measuring effective spreads by matching trades with quotes, which is more sensitive to the quality of the data (see Holden & Jacobsen, 2014).

[Insert Table 1.6 about here.]

1.4.3 Liquidity Premia and the Size Effect

The seminal work by Amihud & Mendelson (1986) presents a theoretical model in which investors demand compensation for holding assets that are more expensive to trade. As a result, expected asset returns increase in the effective bid-ask spread. Due to the difficulties in estimating the bid-ask spread over long time horizons, the literature has considered different proxies of (il)liquidity to study the empirical data (see *e.g.*, Amihud, 2002; Pástor & Stambaugh, 2003; Goyenko et al., 2009; Amihud, 2019, among others).

We illustrate how EDGE can add to this literature by measuring the feature of interest directly and with high accuracy from the perspective of theoretical models (Amihud & Mendelson, 1986).[19] Following the literature that has documented the size, book-to-market, or profitability anomalies (Fama & French, 1992; Novy-Marx, 2013), we run Fama–MacBeth regressions (Fama & MacBeth, 1973) of the monthly excess return on the average spread in the previous year. We include the control variables for size, book-to-market, short-term reversals, momentum, and gross profitability (*e.g.*, Novy-Marx, 2013). We further control for volatility to disentangle liquidity effects from risk (*e.g.*, Amihud, 2002). To construct the control variables, we use the COMPUSTAT database. For this reason, the analysis is carried out in the period 1963–2021, and it includes common shares trading on NYSE, AMEX, and NASDAQ with CRSP share codes of 10 or 11. We require that a firm has appeared on COMPUSTAT for two years to be included (Fama & French, 1993). Financial firms are excluded.

Table 1.7 reports the results. EDGE is significantly priced (t-stat 5.25). As the bid-ask spread has a direct interpretation of transaction costs, we can easily translate the slope coefficients to economic magnitudes. We find that the monthly return premium of an asset with a bid-ask spread of 1.0% is 0.22%.[20] Put differently: Investors demand compensation for transaction costs within a holding period between four and five months (1/0.22).

Columns 2 to 4 report the results for AR, CS, and ROLL, respectively. As these estimators overstate bid-ask spreads where they are expected to be the smallest and understate bid-ask spreads where they are expected to be the largest (see Section 1.4.1), we expect these cross-sectional regression coefficients to be biased. Specifically, as AR and CS are mostly downward biased (Figure 1.7), we expect the regression coefficients to be upward biased. Similarly, as the Roll (1984) estimator is dominated by an upward bias, we expect its regression coefficient to be downward biased. The results in Table 1.7 are consistent with these intuitions. We find that AR overstates the EDGE estimate by 37% (column 2), CS by 111% (column 3), and the Roll (1984) estimator understates the coefficient by 30% (column 4).

Furthermore, columns 5 to 7 show that AR, CS, and ROLL become insignificant after controlling for EDGE. We conclude that EDGE subsumes the explanatory power of earlier

[19] A more thorough examination of the relationship between the bid-ask spread and asset prices can be found in Ardia et al. (2022).

[20] We recall from Fama (1976) that the slope coefficients of the Fama–MacBeth regressions can be interpreted as the *return* to zero-net investment portfolios that have a characteristic of interest of exactly 1.0 (and all other characteristics of exactly zero in the multivariate case).

spread estimators in multivariate regressions, similar to the book-to-market ratio when contested with other value measures in multivariate regressions (Fama & French, 1992), or gross profitability when challenged with other profitability measures (Novy-Marx, 2013).

Finally, we find that also size (t-stat -3.96) becomes insignificant after controlling for EDGE (t-stat -1.50). This finding corroborates the idea that any size effect (Banz, 1981) is, in fact, a consequence of a spread effect, with firm size serving as a proxy for liquidity (see *e.g.*, Amihud & Mendelson, 1986).

[Insert Table 1.7 about here.]

1.5 Conclusion

This paper derives an efficient estimator of the bid-ask spreads from open, high, low, and close prices (EDGE). The estimator is asymptotically unbiased and optimally combines the full set of price data to minimize the estimation variance. EDGE is derived under more general conditions than other estimators and it can be applied to wider cross-sections and longer time-series at any frequency. Our results show that EDGE is always the best choice in all scenarios. Because other estimators overstate bid-ask spreads where they are expected to be the smallest and understate bid-ask spreads where they are expected to be the largest, EDGE should prove useful in a variety of research areas. Specifically, our estimator makes it possible to reliably estimate bid-ask spreads for a variety of assets regardless of their trading frequency, in low and high frequency, for markets that do not report bid and ask data, when quotes are unreliable or expensive to obtain, or when trades cannot be reliably matched with quotes.

We illustrate how EDGE improves inference in a variety of applications. First, in the historical sample 1926–1999, we find that EDGE is often about two times larger than the next best estimator from daily data. This finding suggests that the liquidity benefits of modern financial markets have been underestimated. This finding also suggests that bid-ask spreads have been underestimated even more in several markets where assets are traded less frequently than U.S. stocks.

Second, portfolio sorts differ substantially depending on the spread estimator used. While portfolio sorts are immune to a monotonic bias in the spread estimate, they are not robust to estimation variance. We show that the larger estimation variance of previous estimators leads to substantially more liquid stocks allocated into supposed-to-be-illiquid portfolios and vice versa. Therefore, sorting on EDGE allows for more powerful empirical tests based on portfolio sorts.

Third, we revisit the role of liquidity premia in explaining expected stock returns. We find an annualized liquidity premium of 2.64% for a portfolio with a bid-ask spread of 1.00%. EDGE subsumes the explanatory power of earlier spread estimators in multivariate regressions, similar to the book-to-market ratio when contested with other value measures in multivariate regressions (Fama & French, 1992), or gross profitability when challenged with other profitability measures (Novy-Marx, 2013). In contrast to earlier estimators, EDGE also drives out the size effect of Banz (1981), in line with the hypothesis that market equity is priced in the cross-section of stock returns because it is merely a proxy for transaction costs.

Taken together, our results show that an accurate measure of bid-ask spreads is essential to improve inference in empirical application and EDGE should prove useful in a variety of research areas.

Bibliography

Abdi, F., & Ranaldo, A. (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies*, *30*(12), 4437–4480.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*, 375–410.

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

Amihud, Y. (2019). Illiquidity and stock returns: A revisit. *Critical Finance Review*, *8*, 203–221.

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

Ardia, D., Guidotti, E., & Kroencke, T. A. (2022). Liquidity premia and bid-ask spreads: Evidence from 94 years in the u.s. stock market. *Working Paper*.

Banz, R. W. (1981). The relationship between return and market value of common stocks. *Journal of Financial Economics*, *9*(1), 3–18.

Barclay, M. J., & Smith Jr., C. W. (1988). Corporate payout policy: Cash dividends versus open-market repurchases. *Journal of Financial Economics*, *22*(1), 61–82.

Ben-Rephael, A., Kadan, O., & Wohl, A. (2015). The diminishing liquidity premium. *Journal of Financial and Quantitative Analysis*, *50*(1/2), 197–229.

Chung, K. H., & Charoenwong, C. (1998). Insider trading and the bid-ask spread. *Financial Review*, *33*(3), 1–20.

Chung, K. H., McInish, T. H., Wood, R. A., & Wyhowski, D. J. (1995). Production of information, information asymmetry, and the bid-ask spread: Empirical evidence from analysts' forecasts. *Journal of Banking & Finance*, *19*(6), 1025–1046.

Coller, M., & Yohn, T. L. (1997). Management forecasts and information asymmetry: An examination of bid-ask spreads. *Journal of Accounting Research*, *35*(2), 181–191.

Conroy, R. M., Harris, R. S., & Benet, B. A. (1990). The effects of stock splits on bid-ask spreads. *Journal of Finance*, *45*(4), 1285–1295.

Copeland, T. E., & Galai, D. (1983). Information effects on the bid-ask spread. *Journal of Finance*, *38*(5), 1457–1469.

Corwin, S. A., & Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance*, *67*(2), 719–760.

Detzel, A. L., Novy-Marx, R., & Velikov, M. (2022). Model selection with transaction costs. Working paper, SSRN 3805379.

Drienko, J., Smith, T., & von Reibnitz, A. (2019). A review of the return–illiquidity relationship. *Critical Finance Review*, *8*, 127–171.

Eleswarapu, V., & Reinganum, M. R. (1993). The seasonal behavior of the liquidity premium in asset pricing. *Journal of Financial Economics*, *34*(3), 373–386.

Fama, E. F. (1976). *Foundations Of Finance*. Basic Books.

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

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

Fong, K. Y., Holden, C. W., & Trzcinka, C. A. (2017). What are the best liquidity proxies for global research? *Review of Finance*, *21*(4), 1355–1401.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*(1), 71–100.

Goyenko, R. Y., Holden, C. W., & Trzcinka, C. A. (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics*, *92*, 153–181.

Hagströmer, B. (2021). Bias in the effective bid-ask spread. *Journal of Financial Economics*, *142*(1), 314–337.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

Harris, L., & Amato, A. (2019). Illiquidity and stock returns: Cross-section and time-series effects: A replication. *Critical Finance Review*, *8*, 173–202.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *Journal of Finance*, *64*(3), 1445–1477.

Hasbrouck, J., & Ho, T. S. (1987). Order arrival, quote behavior, and the return-generating process. *Journal of Finance*, *42*(4), 1035–1048.

Holden, C. W., & Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance*, *69*(4), 1747–1785.

Jahan-Parvar, M. R., & Zikes, F. (2019). When do low-frequency measures really measure transaction costs? Working paper, SSRN 3418948.

Johann, T., & Theissen, E. (2017). The best in town: A comparative analysis of low-frequency liquidity estimators. Working paper, SSRN 2905032.

Karnaukh, N., Ranaldo, A., & Söderlind, P. (2015). Understanding FX liquidity. *Review of Financial Studies*, *28*(11), 3073–3108.

Lesmond, D. A., Ogden, J. P., & Trzcinka, C. A. (1999). A new estimate of transaction costs. *Review of Financial Studies*, *12*(5), 1113–1141.

Nieto, B. (2018). Bid–ask spread estimator from high and low daily prices: Practical implementation for corporate bonds. *Journal of Empirical Finance*, *48*, 36–57.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1–28.

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, *29*, 104–147.

Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, *39*(4), 1127–1139.

Tremacoldi-Rossi, P., & Irwin, S. H. (2019). A problem of zeros: The misbehavior of simple bid-ask spread estimators. Working Paper.

Figures

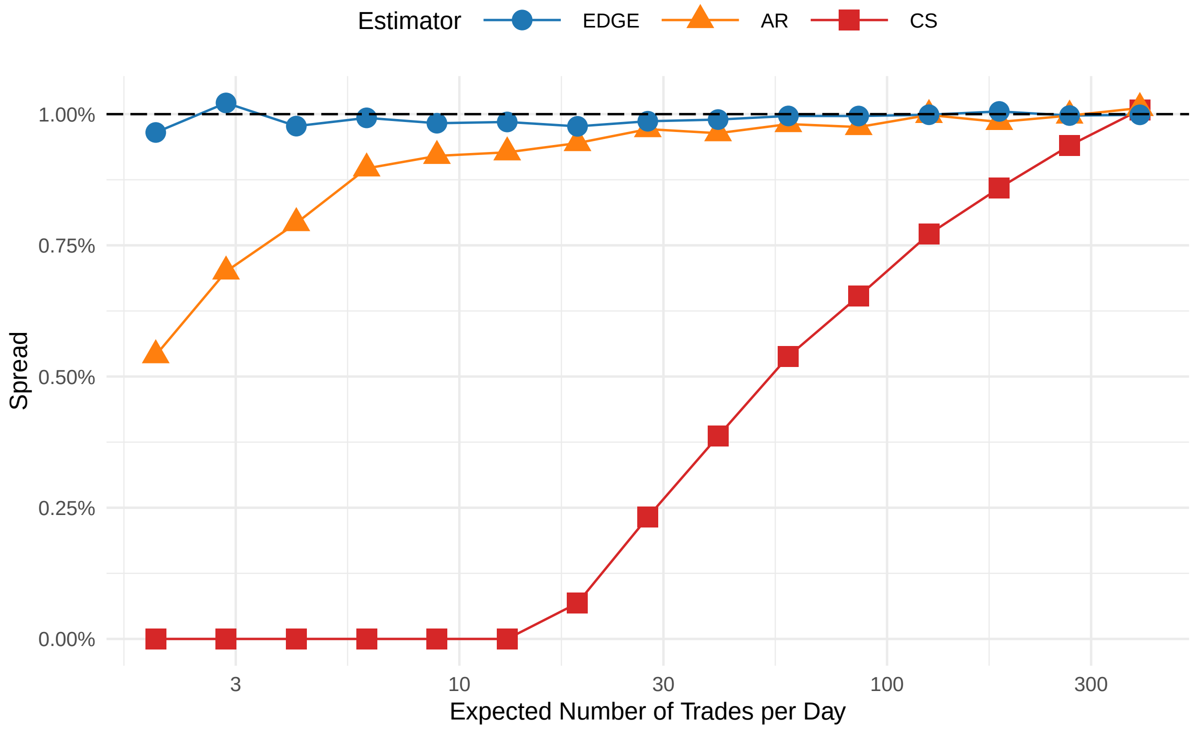


Figure 1.1: Spread estimates (y-axis) in function of the expected number of trades (x-axis). The price process is simulated using the low-frequency setting described in Section 1.2.1 and the simulations use a constant spread of 1%. The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per day is specified by the horizontal axis. EDGE is the estimator proposed in this paper, CS is the estimator in Corwin & Schultz (2012), and AR is the estimator in Abdi & Ranaldo (2017).

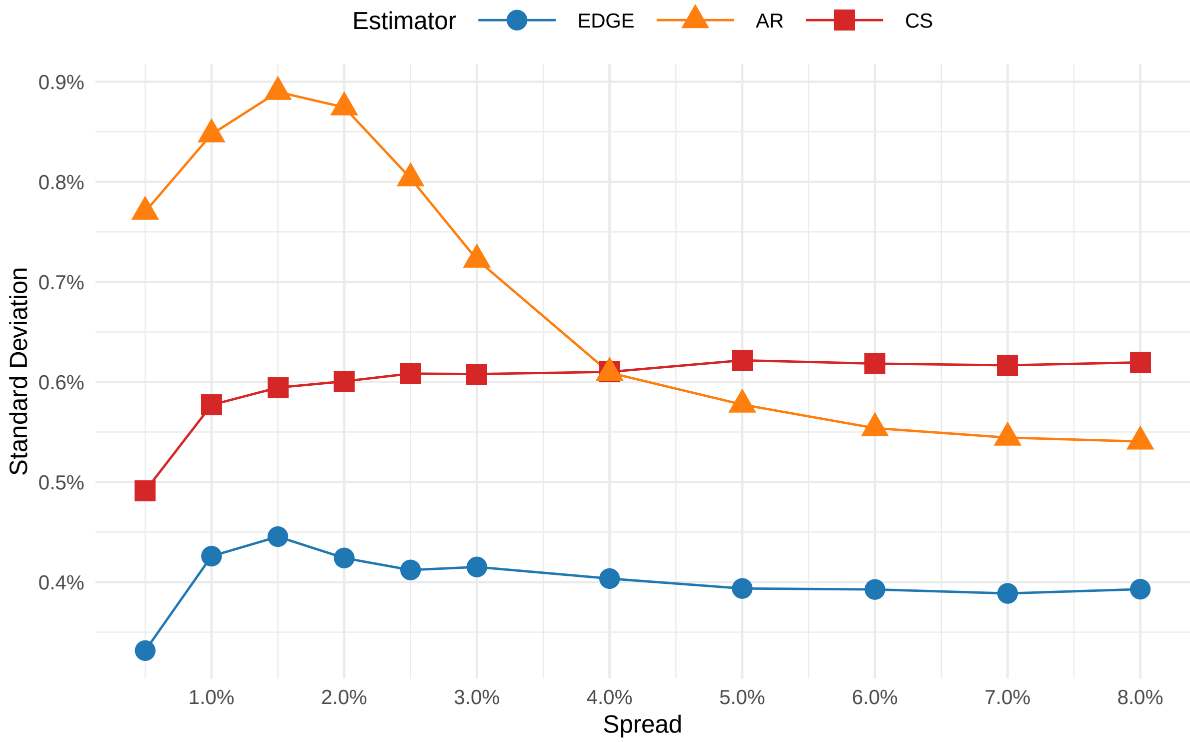


Figure 1.2: Standard deviation of spread estimates (y-axis) for several spread levels used in simulations (x-axis). The price process is simulated using the low-frequency setting described in Section 1.2.1. The simulations use 100% probability to observe a trade each minute (390 trades per day), so that all the estimators are unbiased (see Figure 1.1) and the minimum-variance estimator coincides with the best estimator in the usual root mean squared error sense. EDGE is the estimator proposed in this paper, CS is the estimator in Corwin & Schultz (2012), and AR is the estimator in Abdi & Ranaldo (2017).

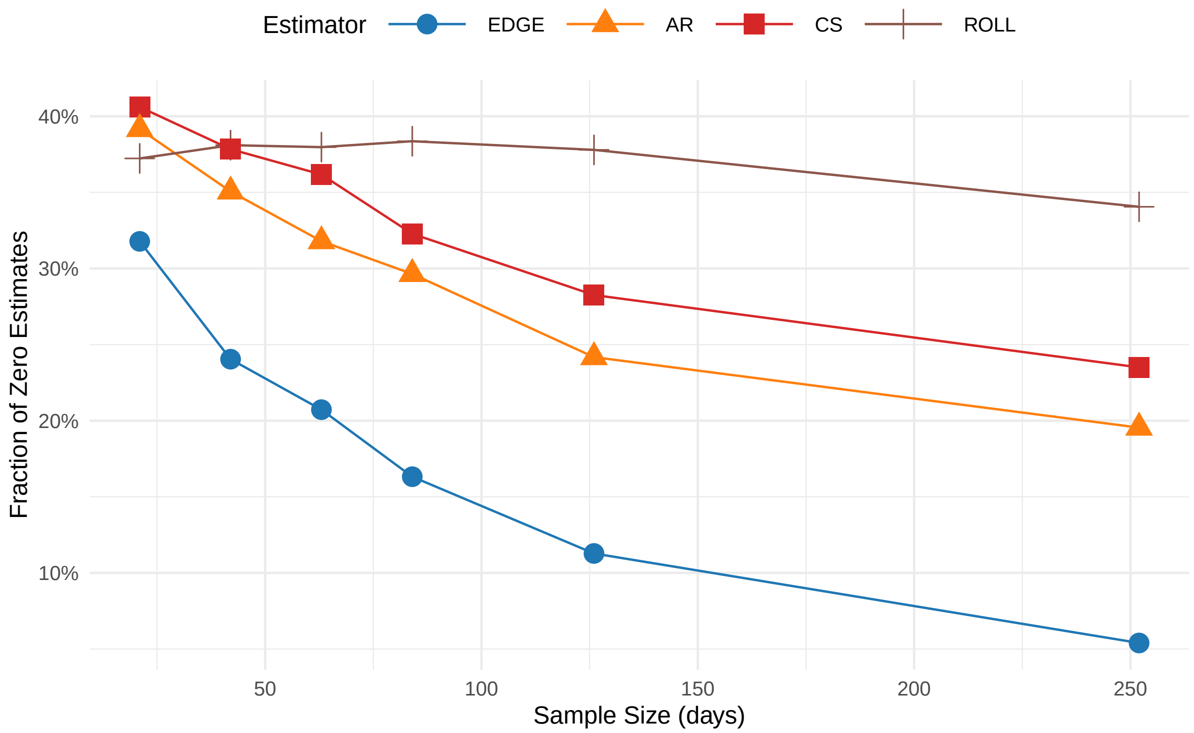


Figure 1.3: Fraction of zero-spread estimates (y-axis) in function of the sample size used for estimation (x-axis), ranging from one month to one year. The price process is simulated using the low-frequency setting described in Section 1.2.1. These simulations use a constant spread of 1%, a 10% probability of observing a trade (for an average of 39 trades per day), and an overnight return normally distributed with mean zero and standard deviation equal to half of the daily volatility. EDGE is the estimator proposed in this paper, CS is the estimator in Corwin & Schultz (2012), and AR is the estimator in Abdi & Ranaldo (2017).

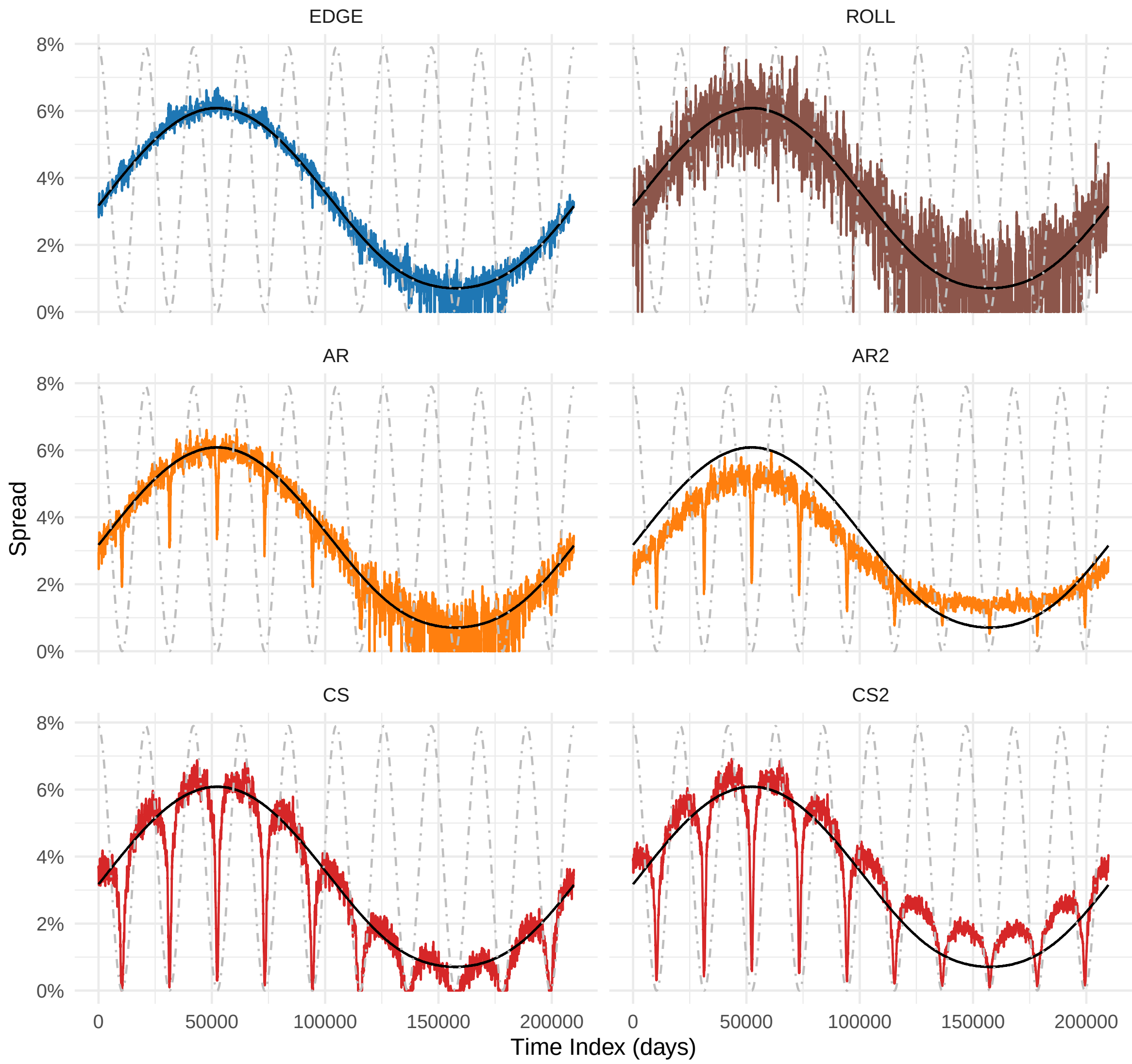


Figure 1.4: Time-series estimates from EDGE as proposed in this paper and those obtained with the estimators in Abdi & Ranaldo (2017) (AR and AR2), Corwin & Schultz (2012) (CS and CS2), and Roll (1984) (ROLL) for a simulated price process. The simulation consists of 10,000 21-day stock-months and each day consists of 390 minutes. For each minute of the day, the true value of the stock price, P_m, is simulated as $P_m = P_{m-1}e^{\sigma x}$, where σ is the standard deviation per minute and x is a random draw from a standard Gaussian distribution. The daily standard deviation equals 3% and the standard deviation per minute equals 3% divided by $\sqrt{390}$. The simulation include an overnight return normally distributed with mean zero and standard deviation equal to half of the daily volatility. The bid (ask) for each minute is defined as P_m multiplied by one minus (plus) half the assumed bid-ask spread. The probability of observing a trade ranges from 0.5% to 99.5% and varies over time according to $p_t = 0.5 + 0.495 \times \cos(\frac{20\pi t}{n})$ where $t = 1, 2, ...$ represents the time index and $n = 10000 \times 21 \times 390$ is the total number of minutes in the simulation. The deterministic component of the spread varies over time according to $\mu_t = 0.03 \times (1 + \sin(\frac{2\pi t}{n}))$. Then, for each minute the spread is randomly drawn from a normal distribution with mean μ_t and standard deviation 0.01. Negative spreads are set to zero. For each day, we use the previous year (21×12 days) to estimate the (root mean square) spread. The solid line in black is the (root mean square) spread in the previous year. The dotted line in grey is the (scaled) average number of trades per day in the previous year.

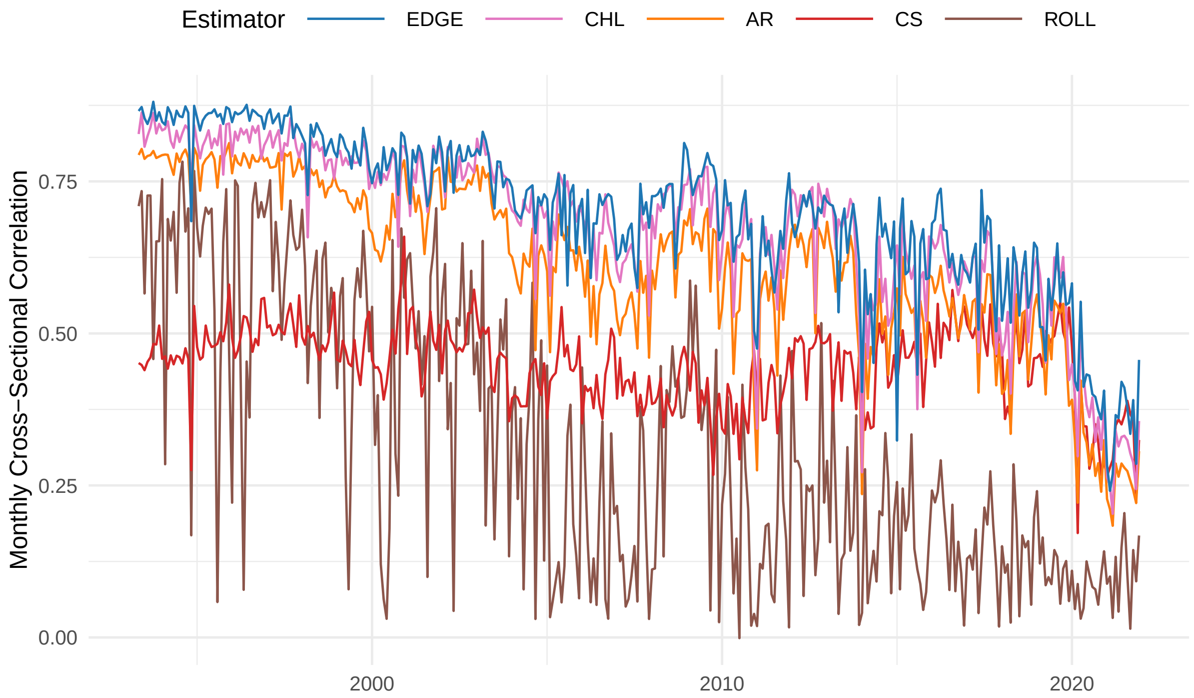


Figure 1.5: Month-by-month cross-sectional correlations with the HJ benchmark (y-axis) in the time period 1993-2021 (x-axis) for various spread estimators as described in Section 1.3.2.

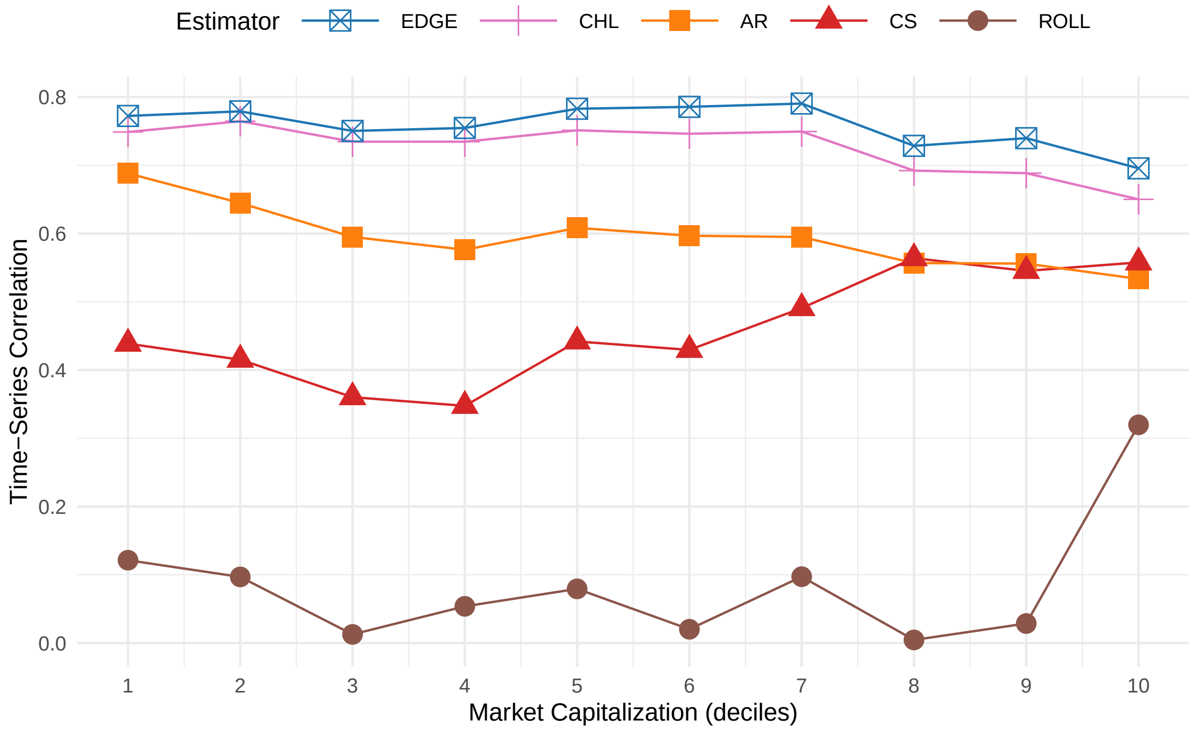


Figure 1.6: Time-series correlations with the HJ benchmark (y-axis) for deciles sorted on size (x-axis, smallest to largest) for various spread estimators as described in Section 1.3.3.

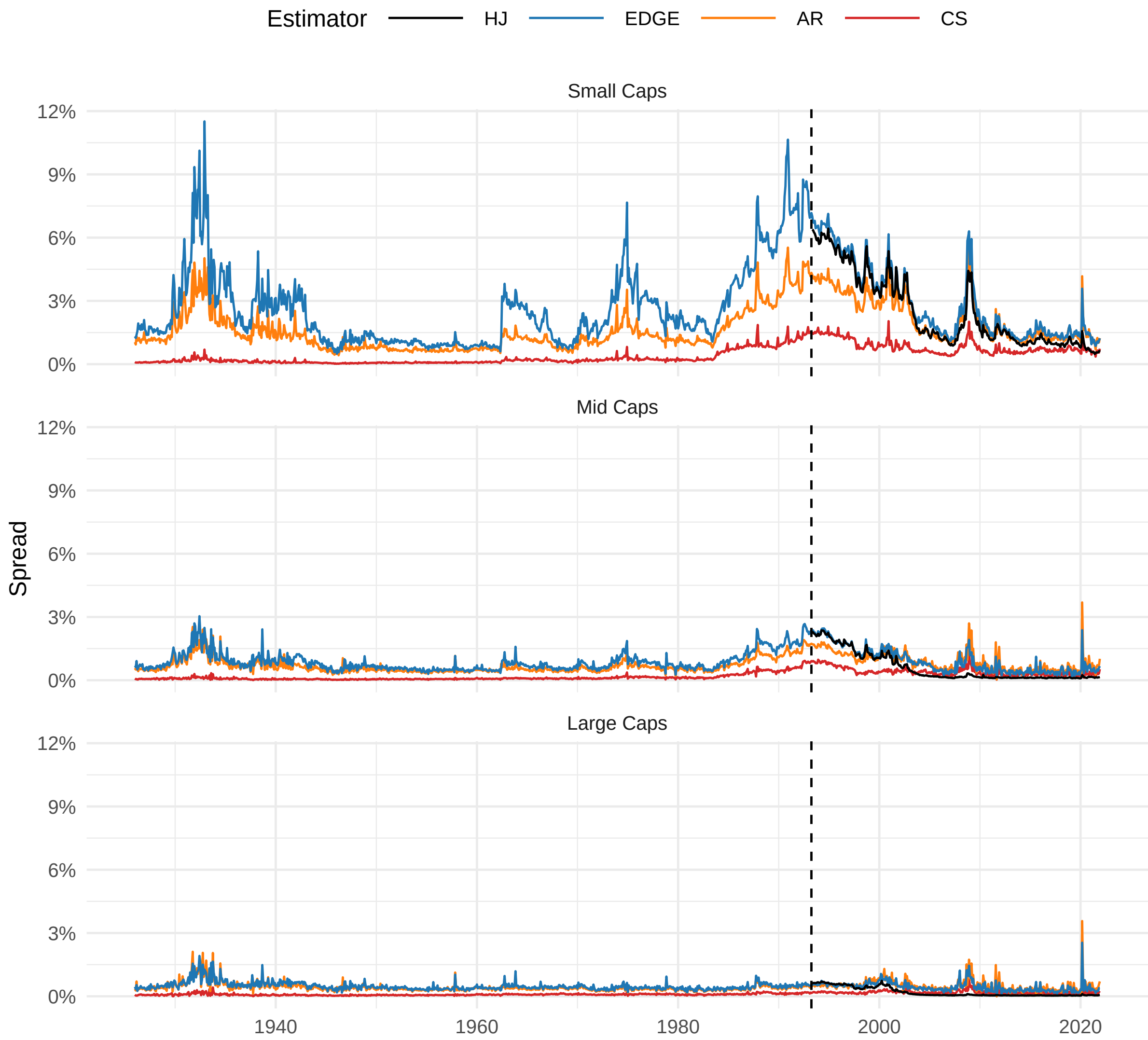


Figure 1.7: Time series of monthly spread estimates for several estimators from daily data (EDGE, AR, CS) and the effective spread benchmark obtained from high-frequency TAQ data (HJ). The panels report the average spreads for (a) small caps, (b) mid caps, and (c) large caps, as described in Section 1.4.1. Averages are trimmed at 2.5%. TAQ data are available since 1993 (vertical line).

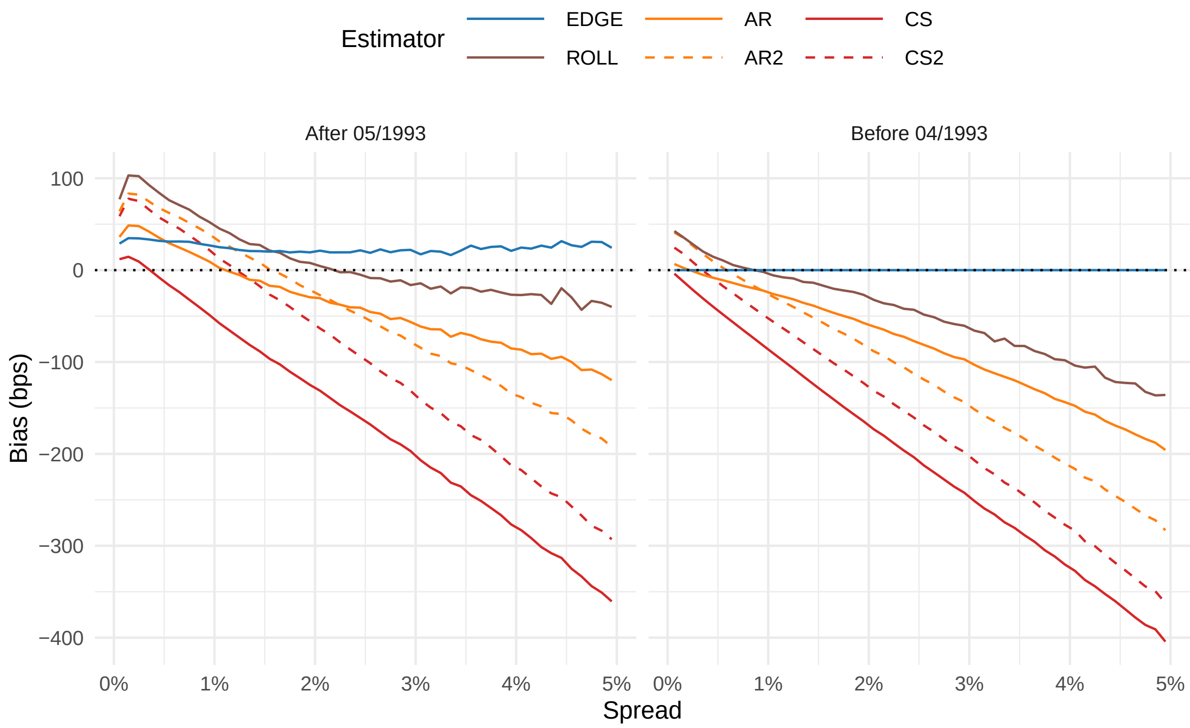


Figure 1.8: Bias in basis points (y-axis) as a function of the benchmark spread (x-axis). Stocks are binned based on the value of the benchmark spread. For each bin and for each estimator, the bias is computed as the average difference between the spread estimate and the benchmark spread. Averages are trimmed at 2.5%. The benchmark spread is HJ after May 1993 (left panel) and EDGE before April 1993 (right panel). The sample period is 1926–2021.

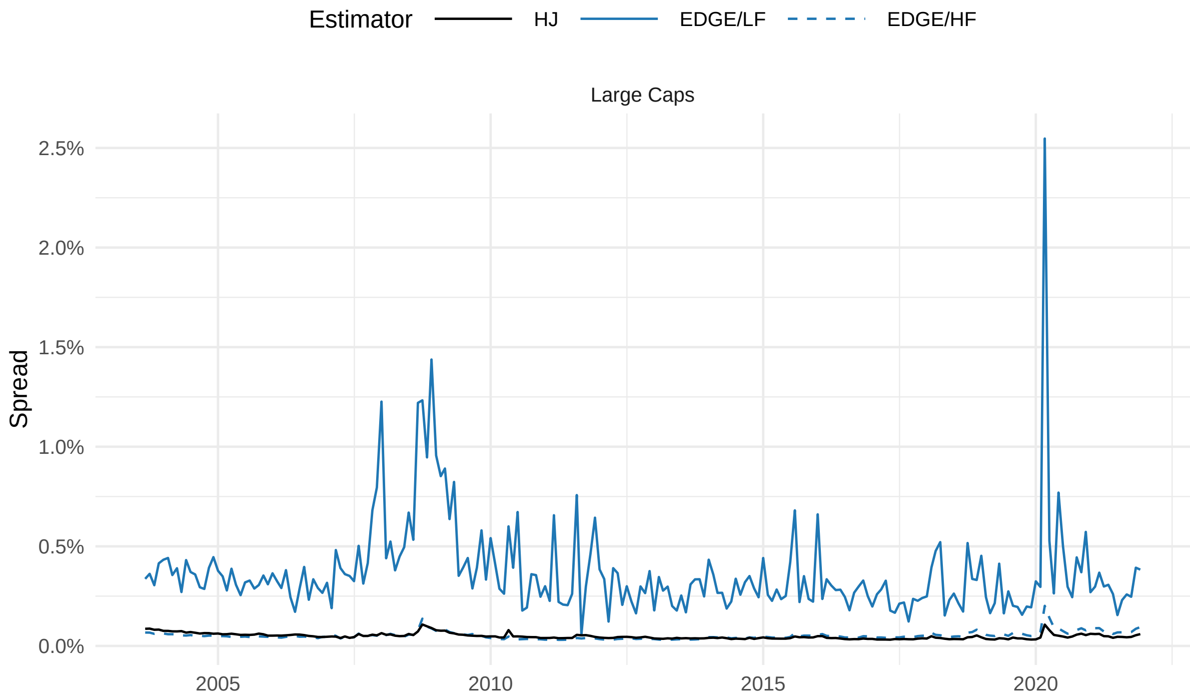


Figure 1.9: Time series of monthly spread estimates for large-cap stocks. The figure reports the average spread within the month. Averages are trimmed at 2.5%. HJ uses the bid-ask spreads computed by matching high-frequency trade and quote data as described in Holden & Jacobsen (2014). EDGE/LF uses EDGE from daily prices. EDGE/HF uses EDGE from minute prices as described in Section 1.4.1.

Tables

Prices	Equations		Prices
O	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[o_t - o_{t-1}, o_{t-1} - o_{t-2}]}{(1-\nu_{o=o})^2}$	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}]}{(1-\nu_{c=c})^2}$	C
OC	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}]}{(1-\nu_{o=c})}$	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}]}{(1-\nu_{o=c=c})(1-\nu_{o=c})}$	CO
OHL	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}]}{(1-k\nu_{o=h,l})}$	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{(1-\nu_{h=l=c})(1-k\nu_{c=h,l})}$	CHL
OHLC	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[\eta_t - o_t, o_t - c_{t-1}]}{(1-k\nu_{o=h,l})}$	$S^2 = -\dfrac{4\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{(1-\nu_{h=l=c})(1-k\nu_{c=h,l})}$	CHLO

Prices	
o,h,l,c	Open, High, Low, Close log-prices.
η	Mid-prices computed as $\eta_t = (l_t + h_t)/2$.

Frequencies	
$\nu_{o=o}$	Fraction of times in which consecutive Open prices match ($o_t = o_{t-1}$).
$\nu_{c=c}$	Fraction of times in which consecutive Close prices match ($c_t = c_{t-1}$).
$\nu_{o=c}$	Fraction of times in which the Open and the Close prices match ($o_t = c_t$).
$\nu_{o=c=c}$	Fraction of times in which both the Close and the Open prices are equal to the previous Close ($o_t = c_t = c_{t-1}$).
$\nu_{h=l=c}$	Fraction of times in which both the High and the Low prices are equal to the previous Close ($h_t = l_t = c_{t-1}$).
$\nu_{o=h,l}$	Computed as $(\nu_{o=h} + \nu_{o=l})/2$, where $\nu_{o=h}$ and $\nu_{o=l}$ are the fraction of times in which the Open price is equal to the High ($o_t = h_t$) or the Low ($o_t = l_t$) price respectively.
$\nu_{c=h,l}$	Computed as $(\nu_{c=h} + \nu_{c=l})/2$, where $\nu_{c=h}$ and $\nu_{c=l}$ are the fraction of times in which the Close price is equal to the High ($c_{t-1} = h_{t-1}$) or the Low ($c_{t-1} = l_{t-1}$) price respectively.

Parameters	
k	Computed as $k = 4b(1-b)$ where b is the probability of the High price to be buyer initiated (or, equivalently, the probability of the Low price to be seller initiated). Appendix 1.A.2 shows that $k = 1$ is a good prior. An estimate is provided in the Appendix and used in Equation (1.12).

Table 1.1: Unbiased bid-ask spread estimators. This table reports bid-ask spread estimators obtained from several combinations of open, high, low, and close prices. The derivation of the estimators is provided in Appendix 1.A.1.

		EDGE	AR	AR2	CS	CS2	ROLL
Panel A: Near-Ideal Conditions (390 Trades per Day)							
Spread 0.50%	Mean	0.44%	0.70%	1.21%	0.60%	1.44%	1.41%
	σ	0.33%	0.77%	0.35%	0.49%	0.33%	1.40%
	$\% \leq 0$	26.69%	46.86%	0.00%	18.51%	0.00%	39.83%
Spread 1.00%	Mean	0.90%	0.95%	1.32%	1.03%	1.75%	1.55%
	σ	0.43%	0.85%	0.37%	0.58%	0.37%	1.46%
	$\% \leq 0$	9.67%	35.08%	0.00%	5.18%	0.00%	36.94%
Spread 3.00%	Mean	2.93%	2.92%	2.41%	2.93%	3.22%	2.88%
	σ	0.41%	0.70%	0.50%	0.61%	0.49%	1.79%
	$\% \leq 0$	0.00%	0.65%	0.00%	0.00%	0.00%	16.76%
Spread 5.00%	Mean	4.96%	4.97%	4.32%	4.90%	4.98%	4.79%
	σ	0.40%	0.58%	0.59%	0.61%	0.57%	2.09%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	6.00%
Spread 8.00%	Mean	7.98%	7.99%	7.58%	7.86%	7.86%	7.75%
	σ	0.38%	0.54%	0.57%	0.62%	0.61%	2.57%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	1.74%
Panel B: With Overnight Returns and 1% Prices Observed ($\approx$ 4 Trades per Day)							
Spread 0.50%	Mean	0.74%	0.70%	1.10%	0.02%	0.35%	1.57%
	σ	0.82%	0.79%	0.35%	0.07%	0.15%	1.55%
	$\% \leq 0$	46.41%	47.64%	0.00%	86.96%	0.00%	39.75%
Spread 1.00%	Mean	0.99%	0.85%	1.19%	0.03%	0.40%	1.70%
	σ	0.91%	0.86%	0.38%	0.09%	0.17%	1.60%
	$\% \leq 0$	36.61%	41.61%	0.00%	82.21%	0.00%	36.93%
Spread 3.00%	Mean	2.88%	2.22%	1.99%	0.27%	0.85%	2.89%
	σ	0.93%	1.04%	0.52%	0.33%	0.30%	1.93%
	$\% \leq 0$	2.58%	9.04%	0.00%	38.85%	0.00%	19.63%
Spread 5.00%	Mean	5.04%	4.02%	3.23%	0.99%	1.56%	4.68%
	σ	0.83%	0.96%	0.71%	0.61%	0.49%	2.24%
	$\% \leq 0$	0.02%	0.56%	0.00%	6.49%	0.00%	8.29%
Spread 8.00%	Mean	8.03%	6.59%	5.33%	2.46%	2.86%	7.53%
	σ	0.85%	1.00%	1.00%	0.95%	0.84%	2.73%
	$\% \leq 0$	0.00%	0.02%	0.00%	0.31%	0.00%	2.65%

Table 1.2: Monthly estimates from simulated daily prices. This table reports monthly spread estimates from this paper (EDGE), Abdi & Ranaldo (2017) (AR and AR2), Corwin & Schultz (2012) (CS and CS2), and Roll (1984) (ROLL) for a simulated price process as described in Section 1.2.2. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of the estimates, and the proportion of estimates that are non-positive across the simulations. Panel B incorporates overnight returns and infrequent observation of prices. In these simulations, we assume a 1% chance of observing a trade at any given minute and overnight returns are normally distributed with mean zero and standard deviation 1.5%.

	EDGE	CHL	AR	AR2	CS	CS2	ROLL
EDGE	–	–	–	–	–	–	–
CHL	87	–	–	–	–	–	–
AR	81	91	–	–	–	–	–
AR2	74	77	90	–	–	–	–
CS	54	53	69	75	–	–	–
CS2	52	48	66	84	90	–	–
ROLL	8	9	12	9	3	3	–

Table 1.3: Correlation of spread estimators. The table reports the overall correlation (in %) of spread estimators. EDGE is the estimator proposed in this paper. CHL is the estimator in Table 1.1 based on close, high, and low prices. AR and AR2 are the estimators in Abdi & Ranaldo (2017). CS and CS2 are the estimators in Corwin & Schultz (2012). ROLL is the estimator in Roll (1984). The sample period is from 1993–2021 (CRSP-TAQ merged sample).

Estimator:	N	Mean	Median	Sd	Cor	MAPE	RMSE	$\% \leq 0$
Units:	1	%	%	%	%	%	1	%
EDGE	1,649,199	2.17	1.02	3.49	**81.82**	**16.50**	**1.24**	**25.06**
CHL	1,649,199	2.28	1.04	4.02	79.45	19.25	1.41	31.91
AR	1,649,199	1.71	0.94	2.54	71.78	19.46	1.42	31.91
AR2	1,649,199	1.68	1.18	1.68	66.94	21.37	1.48	–
CS	1,649,199	0.67	0.28	1.14	52.32	32.04	2.05	29.06
CS2	1,649,199	1.31	0.94	1.31	45.12	31.52	2.30	–
ROLL	1,649,199	2.68	1.35	61.21	3.50	24.53	1.80	32.58
HJ	1,649,199	1.75	0.71	2.50	–	–	–	–

Table 1.4: Summary statistics. The table reports the number of stock-month spread estimates, the mean, median, and standard deviation for the estimates and for the HJ benchmark. EDGE is the estimator proposed in this paper. CHL is the estimator in Table 1.1 based on close, high, and low prices. AR and AR2 are the estimators in Abdi & Ranaldo (2017). CS and CS2 are the estimators in Corwin & Schultz (2012). ROLL is the estimator in Roll (1984). The table further shows the correlation of the positive estimates with the benchmark, the mean absolute percentage error (MAPE) and the root mean squared error (RMSE) computed on the log-spreads (Appendix 1.A.4), and the proportion of spread estimates that are non-positive. The sample period is from 1993–2021 (CRSP-TAQ merged sample).

Group	Spread	Correlation (%) EDGE	CHL	AR	CS	ROLL	%≤ 0 EDGE	CHL	AR	CS	ROLL
Panel A: Analysis across Different Markets											
NYSE	0.15%	**71**	70	55	57	1	**40**	44	44	42	**40**
AMEX	1.71%	75	**76**	67	47	9	**25**	31	31	40	33
NASDAQ	1.30%	**81**	78	70	49	4	**18**	26	26	22	29
Panel B: Analysis across Time Periods											
1993–1996	2.53%	**86**	82	79	49	29	**15**	23	23	26	26
1997–2000	1.70%	**82**	78	73	51	14	**22**	30	30	34	32
2001–2002	1.26%	**78**	77	74	50	13	**24**	31	31	35	32
2003–2007	0.31%	**74**	73	65	45	8	**26**	34	34	29	35
2008–2011	0.25%	**74**	72	65	46	1	**26**	32	32	27	33
2012–2015	0.18%	**61**	**61**	55	45	6	31	37	37	**25**	36
2016–2021	0.18%	**54**	49	42	44	1	34	39	39	**28**	35
Panel C: Analysis across Market Capitalization											
Group 1	2.99%	**78**	76	68	44	12	**15**	23	23	25	26
Group 2	1.77%	**76**	74	59	36	2	**17**	24	24	25	27
Group 3	0.76%	**79**	75	61	44	3	**22**	30	30	26	33
Group 4	0.23%	**77**	73	58	52	1	33	39	39	**32**	37
Group 5	0.08%	**73**	68	55	56	4	38	44	44	**37**	39
Panel D: Analysis across Spread Sizes											
Group 1	0.08%	23	**24**	22	**24**	0	42	46	46	**38**	40
Group 2	0.26%	**49**	43	35	41	3	34	40	40	**33**	39
Group 3	0.73%	**64**	60	47	47	2	**24**	33	33	28	35
Group 4	1.73%	**72**	71	59	45	7	**15**	24	24	24	29
Group 5	4.17%	**76**	74	67	41	3	**10**	16	16	22	19

Table 1.5: Monthly Estimates from Daily Prices. This table reports group-specific correlations of positive spread estimates with the HJ benchmark (Holden & Jacobsen, 2014). The table also reports the fraction of estimates that are non-positive and the median benchmark spread per group. The highest correlation and the lowest fraction of non-positive estimates per group are highlighted in bold. EDGE is the estimator proposed in this paper. CHL is the estimator in Table 1.1 based on close, high, and low prices. AR and AR2 are the estimators in Abdi & Ranaldo (2017). CS and CS2 are the estimators in Corwin & Schultz (2012). ROLL is the estimator in Roll (1984). All estimators are based on daily prices using a monthly estimation window. The sample period is from 1993–2021 (CRSP-TAQ merged sample).

	AR	AR2	CS	CS2	ROLL
−4	0.1%	0.4%	2.2%	2.6%	0.5%
−3	0.7%	1.7%	4.2%	4.3%	2.0%
−2	3.5%	5.4%	8.0%	7.5%	6.3%
−1	16.3%	16.5%	14.8%	14.0%	18.3%
0	58.2%	48.5%	33.9%	33.9%	44.6%
+1	17.4%	22.1%	24.5%	25.7%	20.1%
+2	2.9%	4.3%	10.2%	9.7%	6.2%
+3	0.7%	0.9%	2.0%	2.0%	1.7%
+4	0.1%	0.1%	0.2%	0.3%	0.3%
Lower quintiles	20.6%	24.0%	29.1%	28.4%	27.1%
Higher quintiles	21.2%	27.4%	37.0%	37.7%	28.3%

Table 1.6: Ranking differences. This table shows how spread rankings differ depending on the spread estimator used. For each stock-month and for each estimator, stocks are ranked in quintiles based on their spread estimate. The table reports the difference in rankings obtained when using various estimators instead of EDGE. A positive (negative) rank difference indicates that a stock is ranked higher (lower) compared to the ranking obtained with EDGE. The columns report the distribution of rank differences for each estimator. The rows “Lower quintiles”, and “Higher quintiles” report the sum of all negative and positive rank differences, respectively. The sample period is from January 1926 to December 2021 and EDGE is replaced with CHL (Table 1.1) when open prices are not available. Zero spread estimates are dropped and the monthly estimate is dropped for all the estimators if it is missing for any of them. The sample consists of 1,292,961 stock-month spread estimates for each estimator.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EDGE	21.56				32.26	21.68	23.94	
	[5.25]				[4.27]	[4.99]	[4.70]	
AR		29.72			−13.86			
		[5.92]			[−1.59]			
CS			45.70			4.11		
			[4.16]			[0.38]		
ROLL				15.22			−2.10	
				[5.10]			[−0.59]	
ME	−0.04	−0.05	−0.09	−0.07	−0.03	−0.03	−0.04	−0.11
	[−1.50]	[−2.16]	[−3.45]	[−2.70]	[−1.26]	[−1.32]	[−1.55]	[−3.96]
B/M	0.26	0.27	0.29	0.27	0.25	0.26	0.25	0.27
	[4.85]	[5.04]	[5.42]	[4.94]	[4.76]	[5.06]	[4.80]	[5.02]
STR	−5.24	−5.26	−5.40	−5.36	−5.26	−5.27	−5.26	−5.51
	[−13.58]	[−13.65]	[−14.03]	[−13.88]	[−13.74]	[−13.84]	[−13.74]	[−14.09]
MOM	0.59	0.58	0.58	0.55	0.59	0.59	0.58	0.56
	[3.97]	[3.95]	[3.94]	[3.76]	[4.03]	[4.06]	[3.94]	[3.78]
GP	0.67	0.67	0.69	0.65	0.66	0.68	0.66	0.64
	[6.07]	[6.09]	[6.31]	[5.92]	[5.97]	[6.22]	[5.99]	[5.75]
SDRET	−21.96	−20.73	−16.55	−19.25	−21.88	−21.95	−21.46	−12.73
	[−6.70]	[−6.69]	[−5.63]	[−6.30]	[−6.96]	[−6.91]	[−6.87]	[−4.02]

Table 1.7: Liquidity premia and the size effect. This table reports Fama–MacBeth regressions of monthly excess returns on bid-ask spreads and control variables. The bid-ask spread is the average of the monthly estimates in the previous 12 months. EDGE is replaced with CHL (Table 1.1) when open prices are not available. AR, CS, and ROLL are the estimators in Abdi & Ranaldo (2017), Corwin & Schultz (2012), and Roll (1984), respectively. ME is the natural logarithm of market capitalization in the previous June (Fama & French, 1992). Additional control variables are book-to-market (B/M), short-term reversals (STR), momentum (MOM), and gross profitability (GP) as in Novy-Marx (2013); SDRET is the standard deviation of daily returns in the previous month. Independent variables are winsorized at the 1% level. The sample includes common shares trading on NYSE, AMEX, and NASDAQ, with CRSP share codes of 10 or 11. Firms are not included in the sample until they have appeared on COMPUSTAT for two years (Fama & French, 1993). Financial firms are excluded. The time period covers July 1963 to December 2021. The slope coefficients are multiplied by 100, and t-statistics are given in parentheses. The intercept is included in the regression but omitted from the table.

1.A Appendix

1.A.1 The Generalized Estimators

We recall the law of total covariance or covariance decomposition formula, which is extensively used to derive the results in this section. If X, Y, and W are random variables on the same probability space, and the covariance of X and Y is finite, then:

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[\mathbb{C}\text{ov}[X, Y \mid W]] + \mathbb{C}\text{ov}[\mathbb{E}[X \mid W], \mathbb{E}[Y \mid W]]\,. \tag{1.16}$$

Here we are interested in the case where $X = Z_t$ and $Y = Z_s$ represent bid-ask bounces as defined in Equation (1.2). To compute their covariance, we condition on W equal to:

$$W = \begin{cases} 1 & \text{if } Z_t = Z_s \\ 0 & \text{otherwise} \end{cases} \tag{1.17}$$

That is, W is equal to one if Z_t and Z_s are the same bid-ask bounce and it is equal to zero otherwise. In the first case, the covariance between Z_t and Z_s reduces to the variance of Z, as they are the same random variable. In the second case, their covariance is zero as different bid-ask bounces are independent by assumption.

$$\mathbb{C}\text{ov}[Z_t, Z_s \mid W] = \begin{cases} \mathbb{V}[Z] & \text{for} \quad W = 1 \\ 0 & \text{for} \quad W = 0 \end{cases} \tag{1.18}$$

As Z_t and Z_s have constant mean conditional on W, Equation (1.16) reduces to:

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{C}\text{ov}[Z_t, Z_s] = \mathbb{E}[\mathbb{C}\text{ov}[Z_t, Z_s \mid W]] = \mathbb{V}[Z]\mathbb{P}[Z_t = Z_s] \tag{1.19}$$

where $\mathbb{P}[Z_t = Z_s]$ is the probability that Z_t and Z_s are the same bid-ask bounce.

Buys and Sells Equally Likely

When B is a Bernoulli random variable with probability of success $b = 0.5$, the variance of $Z = S(B - 0.5)$ is:

$$\begin{aligned} \mathbb{V}[Z] &= \mathbb{V}[S(B - 0.5)] \\ &= \mathbb{E}[S^2(B - 0.5)^2] - \mathbb{E}[S(B - 0.5)]^2 \\ &= \mathbb{E}[S^2]\mathbb{E}[B^2 - B + 0.25] - \big(\mathbb{E}[S]\mathbb{E}[B - 0.5]\big)^2 \\ &= \frac{\mathbb{E}[S^2]}{4} - \mathbb{E}[S]^2(b - 0.5)^2 \\ &= \frac{\mathbb{E}[S^2]}{4}\,. \end{aligned} \tag{1.20}$$

In other words, if the spread is not constant, then the variance of Z formally measures the mean squared spread. In case the variance of S is small (*i.e.*, the spread does not vary widely around its mean), we can approximate $\mathbb{E}[S^2] \approx \mathbb{E}[S]^2$ so that $\mathbb{V}[Z]$ becomes, at least approximately, a measure for the average spread. For simplicity, in the paper we write $S^2 = \mathbb{E}[S^2] \approx \mathbb{E}[S]^2$.

Buys and Sells not Equally Likely

If B is a Bernoulli random variable with arbitrary probability of success b, we have:

$$\mathbb{V}[Z] = \frac{\mathbb{E}[S^2]}{4} - \mathbb{E}[S]^2(b-0.5)^2 \approx \mathbb{E}[S]^2 b(1-b)\,. \tag{1.21}$$

C Prices

We need to compute the covariance:

$$\mathbb{C}\mathrm{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = \mathbb{C}\mathrm{ov}[Z_{c_t} - Z_{c_{t-1}}, Z_{c_{t-1}} - Z_{c_{t-2}}]\,, \tag{1.22}$$

where c_t is the closing log-price and Z_{c_t} is the corresponding bid-ask bounce in Equation (1.2). By expanding Equation (1.22), we have:

$$\begin{aligned} \mathbb{C}\mathrm{ov}[Z_{c_t} - Z_{c_{t-1}}, Z_{c_{t-1}} - Z_{c_{t-2}}] = {} & \mathbb{C}\mathrm{ov}[Z_{c_t}, Z_{c_{t-1}}] \\ & + \mathbb{C}\mathrm{ov}[Z_{c_t}, -Z_{c_{t-2}}] \\ & + \mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, Z_{c_{t-1}}] \\ & + \mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, -Z_{c_{t-2}}]\,. \end{aligned} \tag{1.23}$$

According to Equation (1.19), the terms in Equation (1.23) are:

$$\begin{array}{ll} \mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, Z_{c_{t-1}}] & = -\mathbb{V}[Z_c] \\ \mathbb{C}\mathrm{ov}[Z_{c_t}, Z_{c_{t-1}}] & = +\mathbb{V}[Z_c]\mathbb{P}[Z_{c_t} = Z_{c_{t-1}}] \\ \mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, -Z_{c_{t-2}}] & = +\mathbb{V}[Z_c]\mathbb{P}[Z_{c_{t-1}} = Z_{c_{t-2}}] \\ \mathbb{C}\mathrm{ov}[Z_{c_t}, -Z_{c_{t-2}}] & = -\mathbb{V}[Z_c]\mathbb{P}[Z_{c_t} = Z_{c_{t-1}}]\mathbb{P}[Z_{c_{t-1}} = Z_{c_{t-2}}] \end{array} \tag{1.24}$$

where $\mathbb{V}[Z_c]$ is the variance of the closing bid-ask bounces. We estimate the probability that two subsequent closing prices incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{c=c}$) in which the two prices are equal (*e.g.*, there is no trade in the second period):

$$\nu_{c=c} \mathrel{\hat{=}} \mathbb{E}[c_t = c_{t-1}] = \mathbb{P}[Z_{c_t} = Z_{c_{t-1}}] = \mathbb{P}[Z_{c_{t-1}} = Z_{c_{t-2}}]\,. \tag{1.25}$$

By rewriting Equation (1.23) with the terms in (1.24)–(1.25), we have:

$$\mathbb{C}\mathrm{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = -\mathbb{V}[Z_c](1-\nu_{c=c})^2\,. \tag{1.26}$$

Finally, Equation (1.20) sets $\mathbb{V}[Z_c] = S^2/4$:

$$\mathbb{C}\mathrm{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = -\frac{S^2}{4}(1-\nu_{c=c})^2\,. \tag{1.27}$$

O Prices

By replacing c_t, c_{t-1}, c_{t-2} with o_t, o_{t-1}, o_{t-2} in Section 1.A.1, we obtain:

$$\mathbb{C}\mathrm{ov}[o_t - o_{t-1}, o_{t-1} - o_{t-2}] = -\frac{S^2}{4}(1-\nu_{o=o})^2\,. \tag{1.28}$$

CO Prices

We need to compute the covariance:

$$\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}] = \mathbb{C}\text{ov}[Z_{o_t} - Z_{c_{t-1}}, Z_{c_{t-1}} - Z_{o_{t-1}}]\,, \tag{1.29}$$

where o_t is the opening log-price; c_{t-1} is the previous close; and Z_{o_t} and $Z_{c_{t-1}}$ are the corresponding bid-ask bounces in Equation (1.2). By expanding Equation (1.29), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{o,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{o,t-1}] &= \mathbb{C}\text{ov}[Z_{o,t}, Z_{c,t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{o,t}, -Z_{o,t-1}] \\ &+ \mathbb{C}\text{ov}[-Z_{c,t-1}, Z_{c,t-1}] \\ &+ \mathbb{C}\text{ov}[-Z_{c,t-1}, -Z_{o,t-1}]\,.\end{aligned} \tag{1.30}$$

According to Equation (1.19), the terms in Equation (1.30) are:

$$\begin{array}{ll}\mathbb{C}\text{ov}[-Z_{c_{t-1}}, Z_{c_{t-1}}] & = -\mathbb{V}[Z_c] \\ \mathbb{C}\text{ov}[Z_{o_t}, Z_{c_{t-1}}] & = +\mathbb{V}[Z_c]\mathbb{P}[Z_{o_t} = Z_{c_{t-1}}] \\ \mathbb{C}\text{ov}[-Z_{c_{t-1}}, -Z_{o_{t-1}}] & = +\mathbb{V}[Z_c]\mathbb{P}[Z_{c_{t-1}} = Z_{o_{t-1}}] \\ \mathbb{C}\text{ov}[Z_{o_t}, -Z_{o_{t-1}}] & = -\mathbb{V}[Z_c]\mathbb{P}[Z_{o_t} = Z_{c_{t-1}}]\mathbb{P}[Z_{c_{t-1}} = Z_{o_{t-1}}]\end{array} \tag{1.31}$$

where $\mathbb{V}[Z_c]$ is the variance of the closing bid-ask bounces. We estimate the probability that the opening price and the previous close incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{o=c=c}$) in which both the open and the close prices in period t are equal to the closing price in period $t-1$ (*e.g.*, there is no trade in period t):

$$\nu_{o=c=c} \mathrel{\hat{=}} \mathbb{E}[o_t = c_t = c_{t-1}] = \mathbb{P}[Z_{o_t} = Z_{c_{t-1}}]\,. \tag{1.32}$$

Moreover, we estimate the probability that the open and the close prices in period $t-1$ incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{o=c}$) in which the open and the close prices are equal (*e.g.*, there is at most one trade in period $t-1$):

$$\nu_{o=c} \mathrel{\hat{=}} \mathbb{E}[o_{t-1} = c_{t-1}] = \mathbb{P}[Z_{c_{t-1}} = Z_{o_{t-1}}]\,. \tag{1.33}$$

By rewriting Equation (1.30) with the terms in (1.31)–(1.33), we have:

$$\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}] = -\mathbb{V}[Z_c](1 - \nu_{o=c=c})(1 - \nu_{o=c})\,. \tag{1.34}$$

Finally, Equation (1.20) sets $\mathbb{V}[Z_c] = S^2/4$:

$$\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}] = -\frac{S^2}{4}(1 - \nu_{o=c=c})(1 - \nu_{o=c})\,. \tag{1.35}$$

OC Prices

We need to compute the covariance:

$$\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}] = \mathbb{C}\text{ov}[Z_{c_t} - Z_{o_t}, Z_{o_t} - Z_{c_{t-1}}]\,, \tag{1.36}$$

where c_t and o_t are the close and the open log-prices; c_{t-1} is the previous close; and Z_{c_t}, Z_{o_t} and $Z_{c_{t-1}}$ are the corresponding bid-ask bounces in Equation (1.2). By expanding Equation (1.36), we have:

$$\begin{aligned} \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, Z_{o,t} - Z_{c,t-1}] = &\, \mathbb{C}\text{ov}[Z_{c,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[-Z_{o,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, -Z_{c,t-1}]\,. \end{aligned} \tag{1.37}$$

According to Equation (1.19), the first two terms in Equation (1.37) are:

$$\begin{aligned} \mathbb{C}\text{ov}[-Z_{o_t}, Z_{o_t}] &= -\mathbb{V}[Z_o] \\ \mathbb{C}\text{ov}[Z_{c_t}, Z_{o_t}] &= +\mathbb{V}[Z_o]\mathbb{P}[Z_{c_t} = Z_{o_t}]\,, \end{aligned} \tag{1.38}$$

where $\mathbb{V}[Z_o]$ is the variance of the opening bid-ask bounces. The third term is identically zero because (a) if at least one trade is observed for period t, then the left hand side is independent from the right hand side; and (b) if no trade is observed for period t, then $Z_{c_t} - Z_{o_t} = 0$. We estimate the probability that the open and the close prices in period t incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{o=c}$) in which the open and the close prices are equal (*e.g.*, there is at most one trade in period t):

$$\nu_{o=c} \mathrel{\hat{=}} \mathbb{E}[o_t = c_t] = \mathbb{P}[Z_{c_t} = Z_{o_t}]\,. \tag{1.39}$$

By rewriting Equation (1.37) with the terms in (1.38)–(1.39), we have:

$$\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}] = -\mathbb{V}[Z_o](1 - \nu_{o=c})\,. \tag{1.40}$$

Finally, Equation (1.20) sets $\mathbb{V}[Z_o] = S^2/4$:

$$\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}] = -\frac{S^2}{4}(1 - \nu_{o=c})\,. \tag{1.41}$$

CHL Prices

Let us define:

$$\eta_t = \frac{h_t + l_t}{2}\,, \quad Z_{\eta_t} = \frac{Z_{h_t} + Z_{l_t}}{2}\,, \tag{1.42}$$

where h_t and l_t are the high and low log-prices, and Z_{h_t} and Z_{l_t} are the corresponding bid-ask bounces in Equation (1.2). We need to compute the covariance:

$$\mathbb{C}\text{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}] = \mathbb{C}\text{ov}[Z_{\eta_t} - Z_{c_{t-1}}, Z_{c_{t-1}} - Z_{\eta_{t-1}}]\,. \tag{1.43}$$

By expanding Equation (1.43), we have:

$$\begin{aligned} \mathbb{C}\text{ov}[Z_{\eta_t} - Z_{c_{t-1}}, Z_{c_{t-1}} - Z_{\eta_{t-1}}] = &\, \mathbb{C}\text{ov}[Z_{\eta_t}, Z_{c_{t-1}}] \\ &+ \mathbb{C}\text{ov}[Z_{\eta_t}, -Z_{\eta_{t-1}}] \\ &+ \mathbb{C}\text{ov}[-Z_{c_{t-1}}, Z_{c_{t-1}}] \\ &+ \mathbb{C}\text{ov}[-Z_{c_{t-1}}, -Z_{\eta_{t-1}}]\,. \end{aligned} \tag{1.44}$$

According to Equation (1.19), the first three terms in Equation (1.44) are:

$$\begin{aligned}
\mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, Z_{c_{t-1}}] &= -\mathbb{V}[Z_c] \\
\mathbb{C}\mathrm{ov}[Z_{\eta_t}, Z_{c_{t-1}}] &= +\mathbb{V}[Z_c]\mathbb{P}[Z_{\eta_t} = Z_{c_{t-1}}] \\
\mathbb{C}\mathrm{ov}[Z_{\eta_t}, -Z_{\eta_{t-1}}] &= +\mathbb{C}\mathrm{ov}[Z_{c_{t-1}}, -Z_{\eta_{t-1}}]\mathbb{P}[Z_{\eta_t} = Z_{c_{t-1}}]
\end{aligned} \tag{1.45}$$

The last term left to compute is:

$$\mathbb{C}\mathrm{ov}[-Z_{c_{t-1}}, -Z_{\eta_{t-1}}] = \frac{\mathbb{V}[Z_h]\mathbb{P}[Z_{c_{t-1}} = Z_{h_{t-1}}] + \mathbb{V}[Z_l]\mathbb{P}[Z_{c_{t-1}} = Z_{l_{t-1}}]}{2} \tag{1.46}$$

where $\mathbb{V}[Z_h]$ and $\mathbb{V}[Z_l]$ are the variance of the bid-ask bounces incorporated in the high and low prices, respectively. As the high prices are usually buyer initiated and the low prices are usually seller initiated (Corwin & Schultz, 2012), $\mathbb{V}[Z_h]$ and $\mathbb{V}[Z_l]$ are generally different from the variance $\mathbb{V}[Z_c]$ of the closing bid-ask bounces. By Equation (1.21), we have:

$$\mathbb{V}[Z_h] = \mathbb{V}[Z_l] = 4b(1-b)\mathbb{V}[Z_c], \tag{1.47}$$

where b is the probability of the high price to be buyer initiated (or, equivalently, of the low price to be seller initiated). We estimate the probability that the mid price and the previous close incorporate the same bid-ask bounce by counting the fraction of times ($\nu_{h=l=c}$) in which both the high and the low prices at time t are equal to the closing price at time $t-1$ (*e.g.*, there is no trade in period t):

$$\nu_{h=l=c} \mathrel{\hat{=}} \mathbb{E}[h_t = l_t = c_{t-1}] = \mathbb{P}[Z_{\eta_t} = Z_{c_{t-1}}]\,. \tag{1.48}$$

Moreover, we estimate the probability that the closing price incorporates the same bid-ask bounce of the high or low prices by counting the fraction of times in which the closing price matches the high ($\nu_{c=h}$) or low ($\nu_{c=l}$) price (*e.g.*, the close price is selected as the high or low price):

$$\begin{aligned}
\nu_{c=h} &\mathrel{\hat{=}} \mathbb{E}[c_{t-1} = h_{t-1}] &&= \mathbb{P}[Z_{c_{t-1}} = Z_{h_{t-1}}] \\
\nu_{c=l} &\mathrel{\hat{=}} \mathbb{E}[c_{t-1} = l_{t-1}] &&= \mathbb{P}[Z_{c_{t-1}} = Z_{l_{t-1}}]
\end{aligned} \tag{1.49}$$

By rewriting Equation (1.44) with the terms in (1.45)–(1.49) and setting $k = 4b(1-b)$, we have:

$$\mathbb{C}\mathrm{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}] = -\mathbb{V}[Z_c](1 - \nu_{h=l=c})(1 - k(\nu_{c=h} + \nu_{c=l})/2)\,. \tag{1.50}$$

Finally, Equation (1.20) sets $\mathbb{V}[Z_c] = S^2/4$:

$$\mathbb{C}\mathrm{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}] = -\frac{S^2}{4}(1 - \nu_{h=l=c})(1 - k(\nu_{c=h} + \nu_{c=l})/2)\,. \tag{1.51}$$

OHL Prices

Let us define:

$$\eta_t = \frac{h_t + l_t}{2}, \quad Z_{\eta_t} = \frac{Z_{h_t} + Z_{l_t}}{2}\,. \tag{1.52}$$

where h_t and l_t are the high and low log-prices, and Z_{h_t} and Z_{l_t} are the corresponding bid-ask bounces in Equation (1.2). We need to compute the covariance:

$$\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}] = \mathbb{C}\text{ov}[Z_{\eta_t} - Z_{o_t}, Z_{o_t} - Z_{\eta_{t-1}}]\,. \tag{1.53}$$

By expanding Equation (1.53), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{\eta_t} - Z_{o_t}, Z_{o_t} - Z_{\eta_{t-1}}] = {} & \mathbb{C}\text{ov}[Z_{\eta_t}, Z_{o_t}] \\ & + \mathbb{C}\text{ov}[-Z_{o_t}, Z_{o_t}] \\ & + \mathbb{C}\text{ov}[Z_{\eta_t} - Z_{o_t}, -Z_{\eta_{t-1}}]\,.\end{aligned} \tag{1.54}$$

Here the third term is identically zero because (a) if at least one trade is observed for period t, then the left hand side is independent from the right hand side; and (b) if no trade is observed for period t, then $Z_{\eta_t} - Z_{o_t} = 0$. The second term reduces to the variance of the opening bid-ask bounces $-\mathbb{V}[Z_o]$. According to Equation (1.19), the first term is:

$$\mathbb{C}\text{ov}[Z_{\eta_t}, Z_{o_t}] = \frac{\mathbb{V}[Z_h]\mathbb{P}[Z_{o_t} = Z_{h_t}] + \mathbb{V}[Z_l]\mathbb{P}[Z_{o_t} = Z_{l_t}]}{2} \tag{1.55}$$

where $\mathbb{V}[Z_h]$ and $\mathbb{V}[Z_l]$ are the variance of the bid-ask bounces incorporated in the high and low prices, respectively. As the high prices are usually buyer initiated and the low prices are usually seller initiated (Corwin & Schultz, 2012), $\mathbb{V}[Z_h]$ and $\mathbb{V}[Z_l]$ are generally different from the variance $\mathbb{V}[Z_o]$ of the opening bid-ask bounces. By Equation (1.21), we have:

$$\mathbb{V}[Z_h] = \mathbb{V}[Z_l] = 4b(1-b)\mathbb{V}[Z_o], \tag{1.56}$$

where b is the probability of the high price to be buyer initiated (or, equivalently, of the low price to be seller initiated). We estimate the probability that the opening price incorporates the same bid-ask bounce of the high or low prices by counting the fraction of times in which the opening price matches the high ($\nu_{o=h}$) or low ($\nu_{o=l}$) price (*e.g.*, the open price is selected as the high or low price):

$$\begin{aligned}\nu_{o=h} \quad & \hat{=}\; \mathbb{E}[o_t = h_t] \quad = \mathbb{P}[Z_{o_t} = Z_{h_t}] \\ \nu_{o=l} \quad & \hat{=}\; \mathbb{E}[o_t = l_t] \quad = \mathbb{P}[Z_{o_t} = Z_{l_t}]\end{aligned} \tag{1.57}$$

By rewriting Equation (1.54) with the terms in (1.55)–(1.57) and setting $k = 4b(1-b)$, we have:

$$\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}] = -\mathbb{V}[Z_o](1 - k(\nu_{o=h} + \nu_{o=l})/2)\,, \tag{1.58}$$

Finally, Equation (1.20) sets $\mathbb{V}[Z_o] = S^2/4$:

$$\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}] = -\frac{S^2}{4}(1 - k(\nu_{o=h} + \nu_{o=l})/2)\,. \tag{1.59}$$

CHLO Prices

By replacing η_t with o_t in Section 1.A.1, we obtain:

$$\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - \eta_{t-1}] = -\frac{S^2}{4}(1 - \nu_{h=l=c})(1 - k(\nu_{c=h} + \nu_{c=l})/2)\,. \tag{1.60}$$

OHLC Prices

By replacing η_{t-1} with c_{t-1} in Section 1.A.1, we obtain:

$$\mathbb{C}\text{ov}[\eta_t - o_t, o_t - c_{t-1}] = -\frac{S^2}{4}(1 - k(\nu_{o=h} + \nu_{o=l})/2)\,. \tag{1.61}$$

1.A.2 The Efficient Generalized Estimator

Moment Conditions and GMM

We notice that all the estimators in Table 1.1 share the common structure:

$$S^2 = -\frac{4\mathbb{C}\text{ov}[r_{1,t}, r_{2,t}]}{\nu} = -4\frac{\mathbb{E}[r_{1,t}r_{2,t}] - \mathbb{E}[r_{1,t}]\mathbb{E}[r_{2,t}]}{\nu} \approx -\frac{4\mathbb{E}[r_{1,t}r_{2,t}]}{\nu}, \tag{1.62}$$

where $r_{1,t}$ and $r_{2,t}$ are some log-returns and ν represents the adjustment for infrequent trades. The approximation is justified by the fact that the average return at daily or higher frequency should be small compared to the spread.[21] From Equation (1.62), we can rewrite each estimator as a moment condition:

$$\mathbb{E}\left[S^2 + \frac{4r_{1,t}r_{2,t}}{\nu}\right] = 0\,. \tag{1.63}$$

Let us introduce, for each estimator i, the random vector $X_{i,t} = -4r^{(i)}_{1,t}r^{(i)}_{2,t}/\nu^{(i)}$ and the corresponding sample mean $\bar{X}_i \equiv \frac{1}{T}\sum_1^T X_{i,t}$. In this notation, the moment conditions become:

$$\mathbb{E}\left[S^2 - X_{i,t}\right] = 0 \quad \text{for} \quad i = 1, 2, \ldots \tag{1.64}$$

By applying GMM, the efficient estimator is given by:

$$\hat{S}^2 = \arg\min_{S^2} \sum_{ij} (S^2 - \bar{X}_i^\intercal)\Omega_{ij}(S^2 - \bar{X}_j)\,, \tag{1.65}$$

where the weighting matrix is the inverse of the covariance matrix $\Omega = \mathbb{V}[S^2 + X_t]^{-1}$, which simplifies to $\Omega = \mathbb{V}[X_t]^{-1}$ as the variance is translation invariant. Therefore, we have a particular case of GMM where the optimal weighting matrix does not depend on the minimizing variable, and the problem reduces to the minimization of a quadratic form. By differentiating Equation (1.65), setting the derivative equal to zero, and solving for S^2, we obtain:

$$\hat{S}^2 = \frac{\sum_i \bar{X}_i \sum_j \Omega_{ij}}{\sum_{ij}\Omega_{ij}} = \sum_i w_i \bar{X}_i \quad \text{with} \quad \begin{cases} \Omega = \mathbb{V}[X_t]^{-1} \\ w_i = \dfrac{\sum_j \Omega_{ij}}{\sum_{ij}\Omega_{ij}} \end{cases}. \tag{1.66}$$

In principle, we could apply GMM using all the estimators in Table 1.1, that is, eight moment conditions. The estimation of ν may also be casted as additional moment conditions. Although this approach would be asymptotically efficient, it is expected to perform poorly in small samples due to the noise in the estimation of the large covariance matrix. In the next section, we introduce prior knowledge to improve the efficiency in small samples.

Prior Knowledge

Figure 1.10 shows the standard deviation of the estimators for several levels of the spread. We notice that the O (C) estimator is dominated by OC (CO), and OC (CO) is dominated

[21] If it was not the case, a spread of 1% would correspond to a daily average return of approximately the same magnitude, that means an average yearly return higher than 200%. This is not the case for most assets. Moreover, it is always possible to detrend the (log) returns, so that the approximation holds without loss of generality.

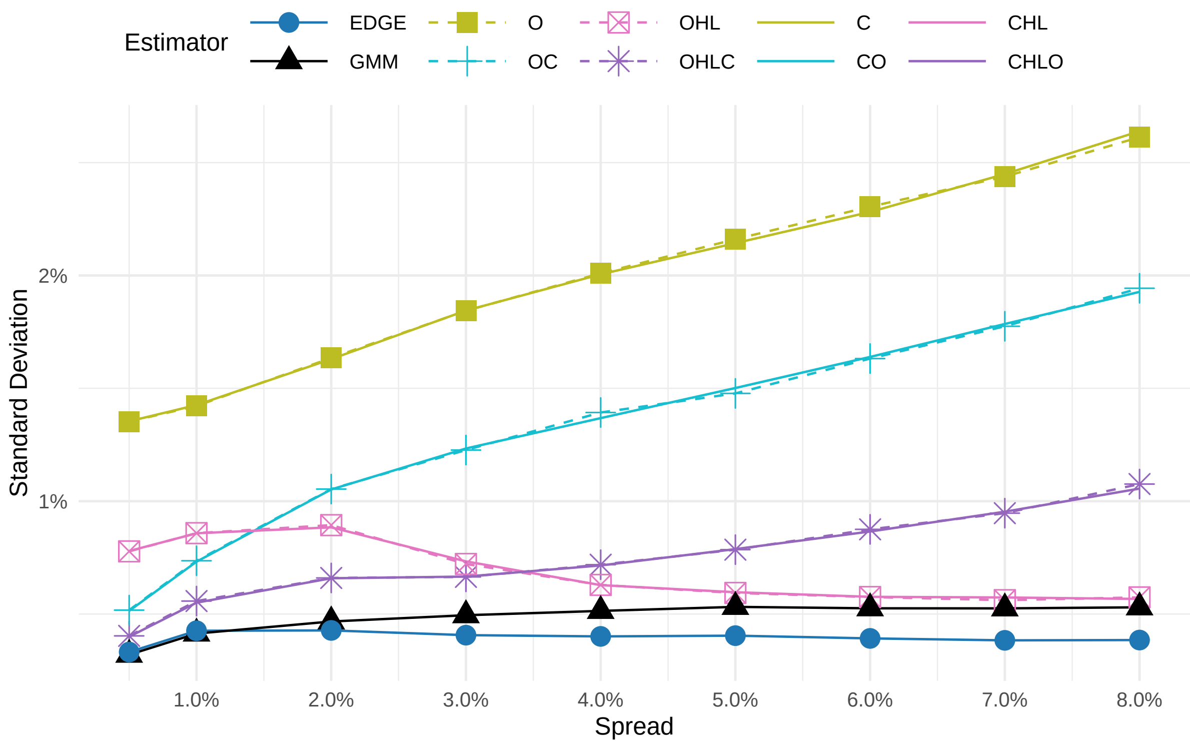


Figure 1.10: Standard deviation of spread estimates (y-axis) for several spread levels used in simulations (x-axis). The price process is simulated using the low-frequency setting described in Section 1.2.1 and the simulations use 100% probability to observe a trade each minute (390 trades per day). The O estimator is affected by the same variance of the C estimator. In the same way, the OC estimator overlaps with CO, OHL with CHL, and OHLC with CHLO. GMM is obtained by computing the pairwise average of the O-C, OC-CO, OHL-CHLO, and OHLC-CHLO estimators, and then applying GMM (1.66) with the corresponding four moment conditions.

by OHLC (CHLO), while OHL (CHL) exhibits a different behaviour. The minimum variance estimators are OHLC and CHLO for small spreads, and OHL and CHL for large spreads.[22]

We consider the moment conditions in Equation (1.63) associated with the four esti-

[22]We highlight that the estimators above sequentially reduce the time interval needed to compute the covariance, thus reducing the sampling error due to the asset's volatility and improving the accuracy of the spread estimates. As such, we expect our OHLC (CHLO) estimator to deliver the most precise estimates of the bid-ask spread. However, when the spread is large compared to the asset's volatility, the OHL (HLC) estimator becomes preferable in practice. As high prices are usually buyer initiated and low prices are usually seller initiated (Corwin & Schultz, 2012), the mid-prices are affected by the spread to a lower extent with respect to the open or close prices. This leads the OHL (CHL) estimator to outperform the OHLC (CHLO) estimator in such cases when the sampling error due to the bid-ask bounces is greater than the sampling error due to the asset's volatility (*e.g.*, in high frequency or highly illiquid markets).

mators:

$$\begin{aligned} \text{OHL}: &\quad \mathbb{E}\big[S^2(1-k\nu_{o=h,l}) + 4(\eta_t - o_t)(o_t - \eta_{t-1})\big] = 0\,, \\ \text{OHLC}: &\quad \mathbb{E}\big[S^2(1-k\nu_{o=h,l}) + 4(\eta_t - o_t)(o_t - c_{t-1})\big] = 0\,, \\ \text{CHL}: &\quad \mathbb{E}\big[S^2(1-\nu_{h=l=c})(1-k\nu_{c=h,l}) + 4(\eta_t - c_{t-1})(c_{t-1} - \eta_{t-1})\big] = 0\,, \\ \text{CHLO}: &\quad \mathbb{E}\big[S^2(1-\nu_{h=l=c})(1-k\nu_{c=h,l}) + 4(o_t - c_{t-1})(c_{t-1} - \eta_{t-1})\big] = 0\,. \end{aligned} \tag{1.67}$$

Then, we notice that the OHL (OHLC) estimator is affected by the same variance of the CHL (CHLO) estimator, as depicted in Figure 1.10.[23] We combine OHL with CHL, and OHLC with CHLO, by summing the corresponding moment conditions. Rearranging the terms gives:

$$\begin{aligned} \text{OHL-CHL}: &\quad \mathbb{E}\left[S^2 + 4\frac{(\eta_t - o_t)(o_t - \eta_{t-1}) + (\eta_t - c_{t-1})(c_{t-1} - \eta_{t-1})}{(1-k\nu_{o=h,l}) + (1-\nu_{h=l=c})(1-k\nu_{c=h,l})}\right] = 0\,, \\ \text{OHLC-CHLO}: &\quad \mathbb{E}\left[S^2 + 4\frac{(\eta_t - o_t)(o_t - c_{t-1}) + (o_t - c_{t-1})(c_{t-1} - \eta_{t-1})}{(1-k\nu_{o=h,l}) + (1-\nu_{h=l=c})(1-k\nu_{c=h,l})}\right] = 0\,. \end{aligned} \tag{1.68}$$

Applying GMM in Equation (1.66) with the two moment conditions above leads to the EDGE in Equations (1.12)–(1.13), with the weights in Equation (1.14).[24] In the next section, we show that k can be set to $k = 1$, and can be refined by setting $k = 4w_1w_2$ with w_1, w_2 given in Equation (1.14). Figure 1.10 shows that, in small samples (*e.g.*, monthly sample of daily data), EDGE achieves a variance lower than the full, asymptotically efficient, GMM.

Estimation of k

We need to estimate $k = 4b(1-b)$ where b is the probability of the high price to be buyer initiated (or, equivalently, the probability of the low price to be seller initiated). To this end, we observe that k affects the estimators in Table 1.1 only through its interaction with ν (specifically, $\nu_{o=h,l}$ or $\nu_{c=h,l}$). Indeed, the spread S is proportional to:

$$\sqrt{\frac{1}{1-k\nu}}\,. \tag{1.69}$$

A precise estimate of k is only needed when the number of trades per period is low, otherwise the adjustment ν is close to zero, and Equation (1.69) is close to one regardless of the value of k. When the number of trades per period is low, b is approximately 50%, that implies $k \approx 1$. A refinement is obtained by observing that if the probability b of

[23]This follows from the functional form of the estimators, that are using (log) returns with the same distribution. For instance, $(\eta_t - o_t)$ in OHL is expected to be distributed as $(c_{t-1} - \eta_{t-1})$ in CHL, and $(o_t - \eta_{t-1})$ as $(\eta_t - c_{t-1})$.

[24]We set the covariance between the two moments equal to zero, as it is expected to be small and hard to estimate in small samples.

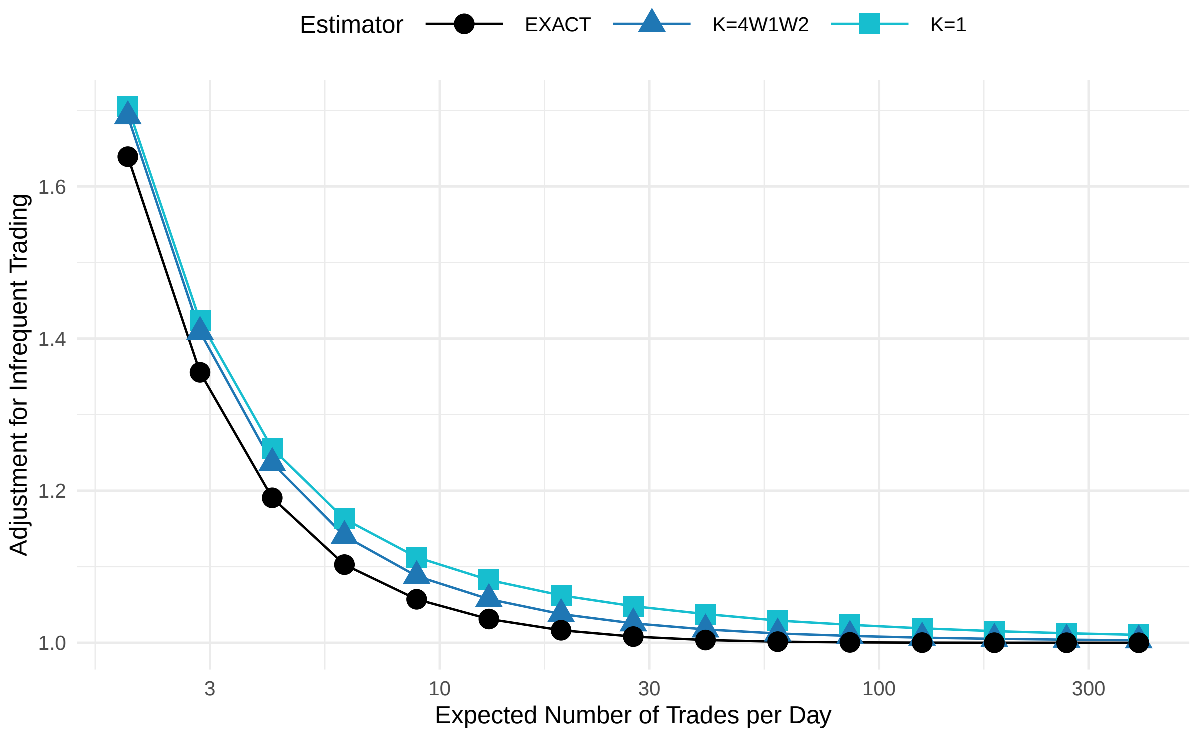


Figure 1.11: The figure displays the value of Equation (1.69) as a function of the number of trades. The price process is simulated using the low-frequency setting described in Section 1.2.1. The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per day is specified by the horizontal axis. The simulations use a constant spread of 1%. Equation (1.69), computed with $k = 1$, or $k = 4w_1w_2$, is compared with the exact values obtained in the simulation.

the high price to be buyer initiated is high (low), then the spread must be large (small) compared to the asset's volatility.[25] In this case, the minimum-variance estimator is OHL-CHL (OHLC-CHLO) as shown in Figure 1.10, and EDGE increases (decreases) the weight w_1 in Equation (1.14). This leads us to identify b with w_1, and we set $k = 4b(1-b) = 4w_1(1-w_1) = 4w_1w_2$.

Figure 1.11 shows the value of Equation (1.69) as a function of the number of trades per period. In this simulations, we compute the exact value of $k = 4b(1-b)$ by using b equal to the fraction of times in which the high price is buyer initiated (rounded up by half the spread). Then, we compute the corresponding value of Equation (1.69), and compare it with the estimates obtained by setting $k = 1$, or $k = 4w_1w_2$.

[25]If the spread is large compared to the asset's volatility, a buyer initiated trade (executed at the ask price) is more likely to be selected as the highest price.

1.A.3 Further Simulation Results

		EDGE	AR	AR2	CS	CS2	ROLL
Panel A: Near-Ideal Conditions (60 Trades per Minute)							
Spread 0.10%	Mean	0.10%	0.10%	0.08%	0.08%	0.10%	0.09%
	σ	0.01%	0.02%	0.01%	0.02%	0.01%	0.06%
	$\% \leq 0$	0.00%	0.60%	0.00%	0.00%	0.00%	16.57%
Spread 0.25%	Mean	0.25%	0.25%	0.22%	0.23%	0.23%	0.25%
	σ	0.01%	0.02%	0.02%	0.02%	0.02%	0.06%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	0.25%
Spread 0.50%	Mean	0.50%	0.50%	0.49%	0.47%	0.47%	0.49%
	σ	0.01%	0.01%	0.01%	0.02%	0.02%	0.08%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Spread 1.00%	Mean	1.00%	1.00%	0.99%	0.97%	0.97%	0.99%
	σ	0.01%	0.01%	0.01%	0.02%	0.02%	0.15%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Panel B: With 2/60 Prices Observed ($\approx$ 2 Trades per Minute)							
Spread 0.10%	Mean	0.09%	0.05%	0.04%	0.00%	0.01%	0.08%
	σ	0.04%	0.03%	0.01%	0.00%	0.00%	0.06%
	$\% \leq 0$	5.65%	12.50%	0.00%	68.01%	0.00%	19.25%
Spread 0.25%	Mean	0.25%	0.14%	0.08%	0.01%	0.02%	0.21%
	σ	0.03%	0.03%	0.02%	0.01%	0.01%	0.06%
	$\% \leq 0$	0.00%	0.05%	0.00%	7.44%	0.00%	1.09%
Spread 0.50%	Mean	0.49%	0.29%	0.17%	0.05%	0.06%	0.43%
	σ	0.05%	0.04%	0.04%	0.02%	0.02%	0.08%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.45%	0.00%	0.00%
Spread 1.00%	Mean	0.98%	0.58%	0.33%	0.13%	0.14%	0.85%
	σ	0.08%	0.08%	0.07%	0.05%	0.05%	0.16%
	$\% \leq 0$	0.00%	0.00%	0.00%	0.05%	0.00%	0.00%

Table 1.8: Hourly estimates from simulated minute prices. This table reports hourly spread estimates from EDGE as proposed in this paper and those obtained with the estimators in Abdi & Ranaldo (2017) (AR and AR2), Corwin & Schultz (2012) (CS and CS2), and Roll (1984) (ROLL) for a simulated price process as described in Section 1.2.2. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of spread estimates, and the proportion of estimates that are non-positive across the simulations. Panel B reports results from simulations incorporating infrequent observation of prices. In these simulations, we assume a 2/60 chance of observing a trade at any given second, for an average of 2 trades per minute.

		EDGE	AR	AR2	CS	CS2	ROLL
Panel A: Near-Ideal Conditions (390 Trades per Day)							
Spread 0.10%	Mean	0.12%	0.34%	1.18%	0.20%	1.23%	0.66%
	σ	0.11%	0.40%	0.10%	0.15%	0.09%	0.75%
	$\%\leq 0$	41.01%	51.44%	0.00%	13.79%	0.00%	48.44%
Spread 0.20%	Mean	0.19%	0.37%	1.19%	0.27%	1.27%	0.68%
	σ	0.13%	0.41%	0.10%	0.16%	0.09%	0.75%
	$\%\leq 0$	24.94%	46.28%	0.00%	5.88%	0.00%	47.36%
Spread 0.30%	Mean	0.26%	0.39%	1.19%	0.35%	1.32%	0.68%
	σ	0.14%	0.41%	0.10%	0.17%	0.09%	0.73%
	$\%\leq 0$	14.15%	44.96%	0.00%	2.40%	0.00%	46.04%
Spread 0.40%	Mean	0.38%	0.45%	1.21%	0.44%	1.38%	0.73%
	σ	0.14%	0.43%	0.10%	0.18%	0.10%	0.76%
	$\%\leq 0$	4.08%	39.09%	0.00%	0.72%	0.00%	44.84%
Spread 0.50%	Mean	0.48%	0.49%	1.22%	0.53%	1.44%	0.77%
	σ	0.13%	0.46%	0.10%	0.18%	0.10%	0.79%
	$\%\leq 0$	1.56%	37.17%	0.00%	0.24%	0.00%	43.76%
Panel B: With Overnight Returns (390 Trades per Day)							
Spread 0.10%	Mean	0.28%	0.39%	1.34%	0.06%	1.18%	0.72%
	σ	0.34%	0.47%	0.12%	0.10%	0.09%	0.83%
	$\%\leq 0$	52.04%	51.80%	0.00%	54.08%	0.00%	48.08%
Spread 0.20%	Mean	0.30%	0.40%	1.35%	0.09%	1.22%	0.70%
	σ	0.33%	0.46%	0.11%	0.11%	0.09%	0.80%
	$\%\leq 0$	47.72%	48.80%	0.00%	40.53%	0.00%	49.28%
Spread 0.30%	Mean	0.33%	0.42%	1.35%	0.13%	1.26%	0.72%
	σ	0.35%	0.48%	0.12%	0.14%	0.09%	0.82%
	$\%\leq 0$	44.84%	49.16%	0.00%	30.82%	0.00%	47.72%
Spread 0.40%	Mean	0.38%	0.46%	1.37%	0.19%	1.31%	0.78%
	σ	0.35%	0.47%	0.11%	0.16%	0.09%	0.82%
	$\%\leq 0$	37.05%	41.61%	0.00%	18.82%	0.00%	44.48%
Spread 0.50%	Mean	0.47%	0.54%	1.38%	0.26%	1.36%	0.83%
	σ	0.37%	0.51%	0.12%	0.17%	0.10%	0.86%
	$\%\leq 0$	29.26%	38.73%	0.00%	11.15%	0.00%	43.65%

Table 1.9: Yearly estimates from simulated daily prices. This table reports yearly spread estimates from EDGE as proposed in this paper and those obtained with the estimators in Abdi & Ranaldo (2017) (AR and AR2), Corwin & Schultz (2012) (CS and CS2), and Roll (1984) (ROLL) for a simulated price process as described in Section 1.2.2. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of the estimates, and the proportion of estimates that are nonpositive across the simulations. Panel B incorporates overnight returns normally distributed with mean zero and standard deviation 1.5%.

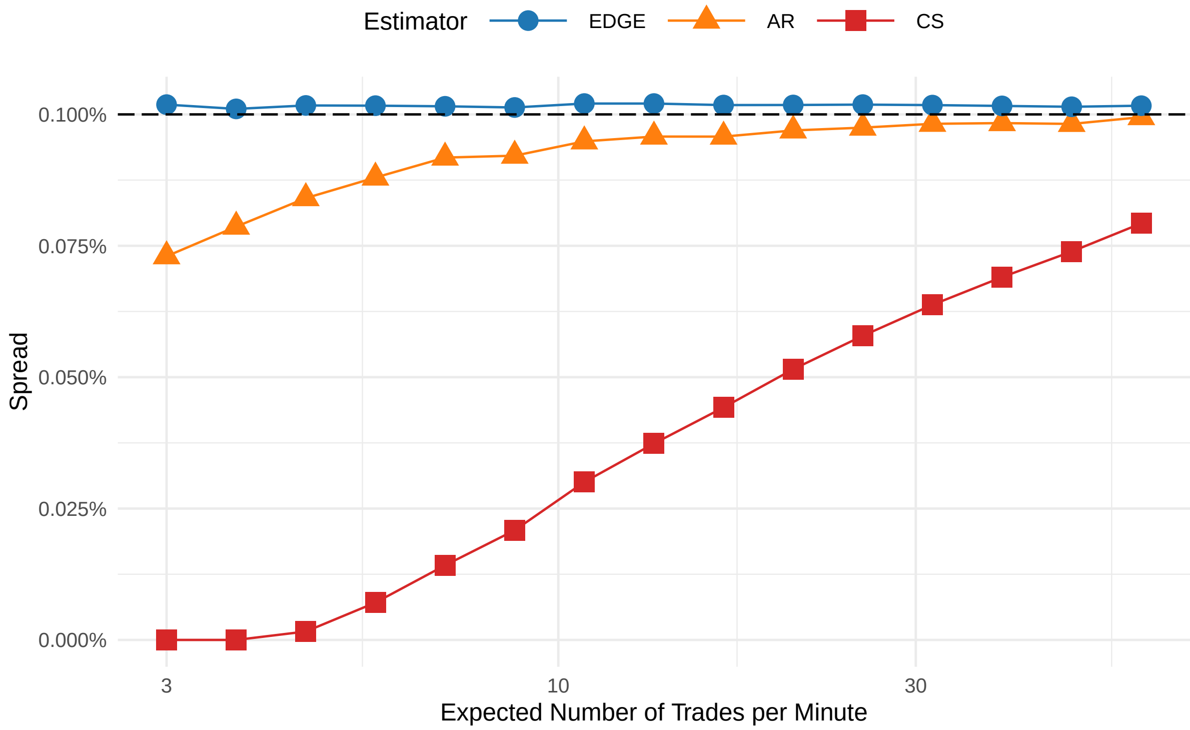


Figure 1.12: High-frequency spread estimates (y-axis) in function of the expected number of trades (x-axis). The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per minute is specified by the horizontal axis. The price process is simulated using the high-frequency setting described in Section 1.2.1 and the simulations use a constant spread of 0.10%. EDGE is the estimator proposed in this paper, CS is the estimator in Corwin & Schultz (2012), and AR is the estimator in Abdi & Ranaldo (2017).

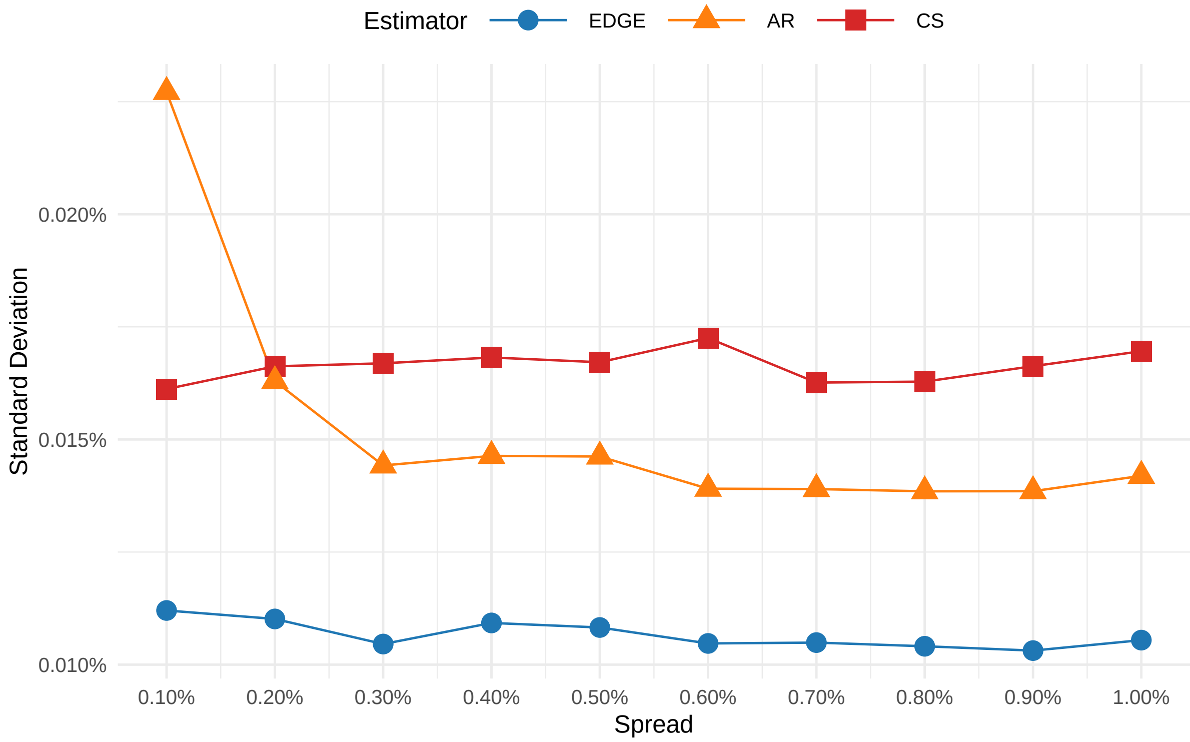
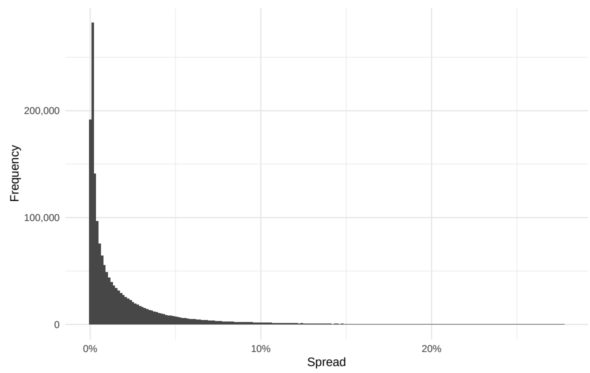
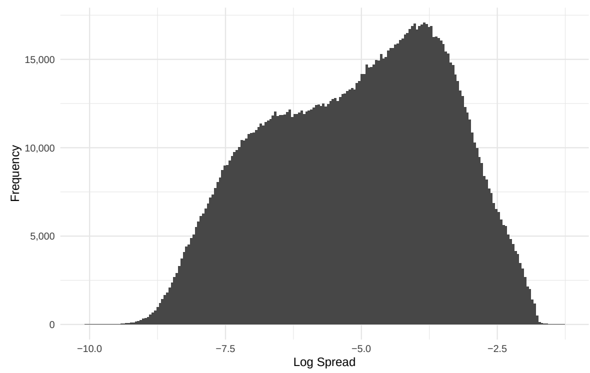


Figure 1.13: Standard deviation of high-frequency spread estimates (y-axis) for several spread levels used in simulations (x-axis). The price process is simulated using the high-frequency setting described in Section 1.2.1. The simulations use 100% probability to observe a trade each second (60 trades per minute), so that all the estimators are unbiased (see Figure 1.12) and the minimum-variance estimator coincides with the best estimator in the usual root mean squared error sense. EDGE is the estimator proposed in this paper, CS is the estimator in Corwin & Schultz (2012), and AR is the estimator in Abdi & Ranaldo (2017).



(a) Spreads



(b) Log-Spreads

Figure 1.14: The histograms show the empirical distribution of effective spreads obtained from TAQ as described in Holden & Jacobsen (2014). Figure (a) reports the distribution of the spreads. Figure (b) reports the distribution of the logarithm of the spreads.

1.A.4 Further Empirical Results

The distribution of effective spreads is highly skewed as displayed in Figure 1.14a. Accordingly, the Mean Absolute Percentage Error (MAPE) can overweight small spreads and the Root Mean Square Error (RMSE) can be severely affected by a few data points in the right tail of the distribution. For this reason, we evaluate the MAPE and RMSE on the logarithmic spreads, which are more symmetrically distributed, as shown in Figure 1.14b.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\log(S_i) - \log(\hat{S}_i)}{\log(S_i)}\right|, \quad \text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(\log(S_i) - \log(\hat{S}_i)\big)^2}. \quad (1.70)$$

	MAPE (%)					RMSE				
	EDGE	CHL	AR	CS	ROLL	EDGE	CHL	AR	CS	ROLL
Panel A: Analysis Across Different Markets										
AMEX	**14**	15	16	45	18	**0.8**	**0.8**	0.9	2.5	1.0
NYSE	**21**	26	25	**21**	33	1.8	2.0	2.0	**1.6**	2.5
NASDAQ	**15**	17	18	35	22	**1.0**	1.2	1.2	2.2	1.5
Panel B: Analysis Across Time Periods										
1993–1996	**9**	11	13	50	14	**0.5**	**0.5**	0.6	2.9	0.7
1997–2000	**12**	15	14	42	18	**0.6**	0.8	0.8	2.3	1.0
2001–2002	**15**	18	18	36	23	**0.9**	1.1	1.1	2.1	1.4
2003–2007	**20**	22	22	24	27	**1.4**	1.6	1.6	1.6	2.0
2008–2011	**23**	27	27	26	33	**1.7**	1.9	1.9	**1.7**	2.4
2012–2015	**20**	23	23	21	30	1.6	1.8	1.8	**1.5**	2.3
2016–2021	22	26	25	**20**	33	1.7	2.0	2.0	**1.5**	2.5
Panel C: Analysis Across Market Capitalization										
Quintile 1	**14**	16	16	49	19	**0.7**	0.8	0.8	2.7	1.0
Quintile 2	**13**	15	16	41	18	**0.8**	0.9	0.9	2.4	1.1
Quintile 3	**14**	17	17	26	22	**1.0**	1.2	1.2	1.7	1.5
Quintile 4	**19**	23	23	21	30	**1.5**	1.7	1.7	**1.5**	2.2
Quintile 5	25	29	28	**20**	36	2.1	2.3	2.3	**1.6**	2.8
Panel D: Analysis Across Spread Sizes										
Quintile 1	26	31	30	**19**	38	2.1	2.4	2.4	**1.6**	2.9
Quintile 2	20	24	24	**19**	31	1.5	1.8	1.7	**1.4**	2.2
Quintile 3	**16**	18	18	21	24	**1.1**	1.2	1.2	1.4	1.6
Quintile 4	**12**	14	14	32	17	**0.7**	0.8	0.8	1.9	1.0
Quintile 5	**12**	14	16	64	16	**0.5**	0.6	0.7	3.2	0.7

Table 1.10: Monthly estimates from daily prices. The table reports the group-specific Mean Absolute Percentage Errors (MAPE) and Root Mean Squared Errors (RMSE) of log-spread estimates with the HJ benchmark as described in Appendix 1.A.4. The lowest MAPE and the lowest RMSE per group are highlighted in bold. EDGE is the estimator proposed in this paper. CHL is the estimator in Table 1.1 based on close, high, and low prices. AR and AR2 are the estimators in Abdi & Ranaldo (2017). CS and CS2 are the estimators in Corwin & Schultz (2012). ROLL is the estimator in Roll (1984). All estimators are based on daily prices using a monthly estimation window. The sample period is from 1993–2021 (CRSP-TAQ merged sample).

Group	Spread	Correlation (%) EDGE	CHL	AR	CS	ROLL	%≤ 0 EDGE	CHL	AR	CS	ROLL
Panel A: Analysis Across Different Markets											
NYSE	0.14%	**74**	**74**	60	65	3	36	41	41	**35**	43
AMEX	1.80%	80	**85**	73	42	53	**19**	24	24	39	35
NASDAQ	1.26%	**91**	90	80	50	52	11	17	17	**9**	26
Panel B: Analysis Across Time Periods											
1993–1996	2.59%	**94**	93	83	45	89	**11**	17	17	23	25
1997–2000	1.83%	**92**	91	84	49	81	**16**	24	24	28	32
2001–2002	1.47%	**89**	**89**	82	47	76	**16**	24	24	30	34
2003–2007	0.35%	87	**88**	83	50	57	21	26	26	**20**	35
2008–2011	0.25%	**85**	84	77	45	6	15	21	21	**12**	28
2012–2015	0.18%	78	**83**	77	54	26	23	28	28	**10**	34
2016–2021	0.18%	**73**	70	63	59	23	29	32	32	**13**	34
Panel C: Analysis Across Market Capitalization											
Group 1	2.81%	87	**88**	76	41	52	**10**	15	15	16	25
Group 2	1.75%	84	**85**	69	34	41	**11**	16	16	16	26
Group 3	0.68%	**87**	86	73	47	32	17	23	23	**16**	33
Group 4	0.20%	**84**	82	70	61	3	28	33	33	**22**	38
Group 5	0.07%	**82**	79	70	68	37	32	39	39	**26**	40
Panel D: Analysis Across Spread Sizes											
Group 1	0.08%	28	28	26	**35**	5	36	42	42	**26**	42
Group 2	0.23%	**56**	51	44	50	6	29	35	35	**24**	40
Group 3	0.66%	**72**	70	58	53	21	**19**	27	27	**19**	37
Group 4	1.64%	78	**80**	68	44	33	**10**	17	17	15	29
Group 5	3.93%	**86**	**86**	73	36	67	**3**	6	6	12	13

Table 1.11: Yearly estimates from daily prices. This table reports group-specific correlations of positive estimates with the HJ benchmark. The table also reports the fraction of spread estimates that are non-positive and the median benchmark spread per group. The highest correlation and the lowest fraction of non-positive estimates per group are highlighted in bold. EDGE is the estimator proposed in this paper. CHL is the estimator in Table 1.1 based on close, high, and low prices. AR and AR2 are the estimators in Abdi & Ranaldo (2017). CS and CS2 are the estimators in Corwin & Schultz (2012). ROLL is the estimator in Roll (1984). All estimators are based on daily prices using a yearly estimation window. The sample period is from 1993–2021 (CRSP-TAQ merged sample).

Chapter 2

Liquidity Premia and Bid-Ask Spreads: Evidence from 94 Years in the U.S. Stock Market

The question of whether less liquid assets demand higher returns has received significant attention in the last decades. The seminal work by Amihud & Mendelson (1986) presents a theoretical model where expected returns increase in the bid-ask spread. However, the difficulty of computing bid-ask spreads over long time periods has led to a variety of liquidity proxies to address this question empirically (*e.g.*, Lesmond et al., 1999; Amihud, 2002; Pástor & Stambaugh, 2003; Hasbrouck, 2009).[1] Hence, there is a gap between the theory of liquidity as measured by bid-ask spreads and the empirical evidence, which adopts different proxies of illiquidity.

Here we fill this gap by using a recently developed measure of the effective bid-ask spread. Ardia et al. (2021) provide an efficient estimator of the bid-ask spread (EDGE) derived from open, high, low, and close prices. The estimator is asymptotically unbiased and significantly more accurate than earlier estimators (Roll, 1984; Corwin & Schultz, 2012; Abdi & Ranaldo, 2017). In this paper, we rely on the NYSE Trade and Quote (TAQ) database to compute the effective spread from high-frequency data in the period 1993-2021. When TAQ data are not available, we use the CRSP database to estimate effective bid-ask spreads from daily prices using EDGE. This allows us to study liquidity as measured by the bid-ask spread since 1927 for common shares listed on NYSE, AMEX, and NASDAQ.

We start by illustrating the difference between our measure of bid-ask spreads and other proxies of illiquidity, and argue that it is important for many empirical applications to use an accurate measure.

Next, we run Fama & MacBeth (1973)-regressions to investigate the relationship between stock returns and the bid-ask spread. We refer to the coefficient on the spread as the "price of illiquidity." The spread itself captures the "amount of illiquidity." The product between the price and the amount of illiquidity is the "illiquidity premium."

We find that the price of illiquidity is 0.22 and it is highly significant. Holding all other characteristics equal, a stock with a spread of 1.00% comes with a 0.22% larger monthly return compared to an ideal zero-spread asset. This number implies that investors expect to be compensated for transaction costs within about 4.5 months (1/0.22). The coefficient is robust across different specifications for the control variables and it remains stable across different sub-samples. When looking at seasonality, we find that the price of illiquidity spikes in January (0.90) but it remains significant also in non-January months (0.15). To study how the price of illiquidity varies across firm size, we split stocks in

[1] See Goyenko et al. (2009) for a survey of liquidity measures.

size groups based on their market capitalization. When we use three size groups, we find that the spread is consistently priced across all groups. When we increase the number of groups to 25, we find that estimates come with larger dispersion the more we move from small-size groups to large-size groups. This suggests that liquidity is priced across all stocks, although samples restricted to larger stocks lack power to detect a significant effect.

We study the illiquidity premium in percentage points and in real dollars. Over the full sample, we find that the average premium (equally weighted) is 0.32% per month. For the market portfolio (value weighted), we estimate a premium of 0.09% per month. The lower premium for the market reflects the fact that stocks with a large market capitalization tend to have smaller spreads (Roll, 1984). The premium of the market portfolio in annual terms is 1.1%. The equity premium is usually estimated to be around 6% (Kroencke, 2017). Accordingly, we can attribute an economically significant share of about 20% of the historical equity premium to compensation for illiquidity.

In real dollars, the equally-weighted premium is 1.5 million per stock per month and it increases to 6.8 millions for the value-weighted average (*i.e.*, for larger stocks). This result illustrates that the lower percentage premium of large-cap stocks actually adds up to a larger dollar amount when moving from percentage returns to total compensation in dollars. For the whole market, we estimate a premium of 4.12 billion dollars per month to compensate for illiquidity.

Next, we empirically assess the predictions in Amihud & Mendelson (1986). To this end, we split stocks in groups based on the their effective bid-ask spread and we study how the price of illiquidity and the illiquidity premium vary across spread size. In line with the predictions, we find that the price of illiquidity declines as we move to higher-spread groups, while the premium increases. Against the predictions, we find that the spread-return relationship is convex.

Finally, we formally show that the Amihud (2002) illiquidity measure (*ILLIQ*) can be seen as a non-linear proxy of the bid-ask spread and that cross-sectional sorts based on *ILLIQ* are equivalent to cross-sectional sorts based on bid-ask spreads. We then illustrate how the "illiquid minus liquid" factor (Amihud et al., 2015) can be closely replicated by replacing *ILLIQ* with bid-ask spreads. In other words, the factor is interpreted as an "expensive minus cheap" (EMC) factor, which goes long in stocks that are expensive to trade (*i.e.*, large bid-ask spreads) and goes short in stocks that are cheap to trade (*i.e.*, small bid-ask spread). This interpretation may apply more generally to applications that rely on cross-sectional sorts on the Amihud (2002) illiquidty measure.

Our results contribute to a large body of literature on liquidity and asset prices. First, our estimates of the price of illiquidity differ substantially from the estimates obtained by Hasbrouck (2009), who finds that "the gross expected monthly return impounds triple or quadruple the effective cost" and this effect is "too large to be consistent with straightforward trading explanations" as it would imply that investors are compensated for transaction costs within about one week. Our results are more in line with the estimate of 0.21 for the price of liquidity originally reported in the sample period 1961–1980 by Amihud & Mendelson (1986).

Second, earlier research shows that liquidity is more strongly priced in January months, but whether the spread is priced also in non-January months is still an open question.

Eleswarapu & Reinganum (1993) find that the spread is reliably priced only during the month of January. Eleswarapu (1997) finds that the spread is more strongly priced in January, but it is also significant in non-January months. Hasbrouck (2009) finds that the coefficients on January effective cost are positive and significant, while those on the non-January components become marginally significant. Recently, Lou & Shu (2017) find that bid-ask spreads are priced only in January, but not in the full sample period. Our results suggest that the bid-ask spread is actually priced throughout the year, but the smaller magnitude in non-January months makes it harder to detect a statistically significant effect.

Third, earlier research documents a significant downward time trend in the price of illiquidity (Ben-Rephael et al., 2015; Drienko et al., 2019; Harris & Amato, 2019). In stark contrast to previous research, our estimates over 10-year rolling windows oscillate symmetrically around the full sample mean. The 90% confidence intervals usually do not include zero, and almost always include the full-sample estimate of 0.22. In the most recent sub-samples, the point estimate is even above the full sample estimate.

Furthermore, our results are in contrast with Ben-Rephael et al. (2015), who report a significant liquidity premium only among the smallest size tercile and conclude that the liquidity premium (p.214) "is narrowly concentrated among small stocks". Instead, our results suggest that the illiquidity is priced across all stocks, although samples restricted to larger stocks lack power to detect a significant effect and make it in general more difficult to find a positive result.

This work provides strong empirical support for the predictions in Amihud & Mendelson (1986). Specifically, our results are in line with the idea that assets with higher spreads are allocated to investors with longer holding periods and that the premium is an increasing function of the spread. However, the data do not support the concave spread-return relationship predicted by the model and we document a convex relation driven by high-spread assets. In other words, the premium of highly illiquid assets is too high to be explained solely in terms of compensation from transaction costs.

Finally, we show that the Amihud (2002) illiquidity measure can be seen as a non-linear proxy for bid-ask spreads. This link between *ILLIQ* and bid-ask spreads has important implications. First, our result is in line with the idea that $\sqrt{ILLIQ}$ should be preferred over *ILLIQ* in empirical studies (see *e.g.*, Chordia et al., 2009; Chelley-Steeley et al., 2015). Second, the return premium associated with *ILLIQ* may be interpreted as compensation for bid-ask spreads, rather than for general illiquidity. This finding contributes to the financial literature that studies why the Amihud (2002) measure is empirically priced (see *e.g.*, Brennan et al., 2013; Lou & Shu, 2017) and which dimensions of liquidity are relevant for stock returns.

The paper is organized as follows. Section 2.1 describes the data and the sample construction. Section 2.2 presents the empirical methodology. Sections 2.3, Section 2.4, and Section 2.5 present the results on the amount of illiquidity, price of illiquidity, and illiquidity premium, respectively. Section 2.6 tests the predictions in Amihud & Mendelson (1986). Section 2.7 shows how the illiquid-minus-liquid factor (Amihud et al., 2015) can be closely replicated by replacing the Amihud (2002) illiquidity measure with bid-ask spreads. Finally, Section 2.8 concludes.

2.1 Data

Our sample consists of NYSE, AMEX, and NASDAQ stocks with CRSP share codes of 10 or 11 (*i.e.*, US common shares) in the period 1927–2021. Returns are adjusted for selection bias using delisting returns and the imputation of Shumway (1997). Excess returns are computed with respect to the risk free rate downloaded from the website of Kenneth R. French.[2] To measure inflation, we use the consumer price index (*CPI*) provided by the Federal Reserve Bank of St. Louis.[3]

Effective bid-ask spreads are derived from high-frequency data using the Trades and Quotes (TAQ) database in the period 1993-2021. Following the recommendations in Holden & Jacobsen (2014), we mainly rely on the Daily TAQ database and we use Monthly TAQ only when Daily TAQ is not available. We download the effective spreads for each stock-day from the Wharton Research Data Services (WRDS) Intraday Indicators using Monthly TAQ for 1993–2003 and Daily TAQ from 2004 onward. We match CRSP and TAQ data using CUSIP identifiers and tickers.[4] Our identification strategy allows us to match about 99% of the stocks in CRSP. The monthly spread for each stock is computed as the average of the daily spread estimates within the month.

To compute effective bid-ask spreads when high-frequency data are not available, we rely on the efficient estimator (EDGE) proposed by Ardia et al. (2021). The estimator allows us to compute bid-ask spreads for each stock-month using daily open, high, low, and close prices, which are directly available from CRSP since 1927. EDGE has been shown to be more accurate and less biased than alternative methods (Roll, 1984; Corwin & Schultz, 2012; Abdi & Ranaldo, 2017). When open prices are missing and EDGE cannot be computed, we use the CHL estimator (Ardia et al., 2021, Table 1).[5]

The main control variables are derived from the CRSP database. We use the market capitalization (*ME*) as of previous June to capture the size effect. Short-term reversal (*STR*) is the return in the previous month. Momentum (*MOM*) is the cumulated return in the previous year, excluding the very last month. Volatility is computed as the standard deviation of daily returns in the previous month (*SDRET*), where we exclude zero-returns from the calculation as they are typically due to absence of trading.

We also include control variables based on COMPUSTAT: the book-to-market ratio (*BE/ME*, Fama & French (1992)) and gross profitability (*GP*, Novy-Marx (2013)). Accounting variables from COMPUSTAT are merged with CRSP starting from 1963 as described in Fama & French (1992). As standard practice, we require firms to have appeared on COMPUSTAT for at least two years (Fama & French, 1993) to be included. Since the accounting-based control variables restrict the sample to the period after 1963, we mainly use them for robustness tests and we do not include them in the baseline results.

[2] `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`

[3] `https://fred.stlouisfed.org/series/CPIAUCNS`

[4] We reconstruct the time series of CUSIPs for each KYPERMNO in CRSP. Similarly, we reconstruct the time series of TICKERs for each KYPERMNO in CRSP. Then, we compute the time series of CUSIPs for each SYMBOL in TAQ using the Monthy TAQ Master files for 1993–2009 and the Daily TAQ Master files in 2010–2021. Finally, we merge the (daily) datasets by matching observations with the same date, with the same CUSIP, and where the TAQ's SYMBOL is equal to the TICKER in CRSP.

[5] Open prices are missing in CRSP from July 1962 through June 1992.

Finally, we study the relation between bid-ask spreads and the Amihud (2002) illiquidity measure. *ILLIQ* is computed for each stock-month by averaging the ratio between absolute daily returns and traded volume. Volume is computed by multiplying the number of shares traded in the day times the closing price. Daily observations with non-positive price or volume are dropped from the calculations.

All dependent variables (*i.e.*, other than excess returns) are winsorized at the 1% level. Table 2.1 reports summary statistics. Our sample consists of about 3.1 million stock-month spread estimates. The number of observations for the bid-ask spread is similar to that of *ILLIQ*. In our regressions, we drop observations with incomplete data. That is, we drop an observation if any of the control variables is missing. However, this is rarely the case, as the main control variables have an extensive data coverage (3.5 million observations). The coverage of the accounting-based variables is more limited (2.4 million observations) and we don't include them in the main results as already discussed.

[Insert Table 2.1 about here.]

2.2 Methodology and Terminology

In this paper, we rely on Fama & MacBeth (1973)-regressions to investigate the relationship between stock returns and the bid-ask spread. For each month t, we run a cross-sectional regression of the type:

$$R_{i,t} - R_{F,t} = a_t + \lambda_t SPREAD_{i,t-1} + \mathbf{b}_t'\mathbf{X}_{i,t-1} + \varepsilon_{i,t}\,, \tag{2.1}$$

where a_t, λ_t, and $\mathbf{b_t}$ are time-varying regression coefficients and the Fama-MacBeth estimates a, λ, and $\mathbf{b}$ are obtained by averaging the coefficients through time. $R_{i,t} - R_{F,t}$ is the excess return over the risk-free rate of stock i in month t, $SPREAD_{i,t-1}$ is the bid-ask spread of stock i measured from past information, and λ_t is the slope coefficient on the spread in month t. The term $\mathbf{b}_t'\mathbf{X}_{i,t-1}$ captures a number of control variables with their respective slope coefficients. The regressions are run on individual stocks rather than portfolios. Individual stocks have a number of advantages over portfolio sorts and they do not depend on the (arbitrary) choice of the test portfolios. On the other hand, individual stocks are more prone to errors-in-variables bias (see *e.g.*, Jegadeesh et al., 2019). That is, the slope coefficient is biased towards zero when the regressor is measured with error. We can substantially mitigate this concern because we use the most accurate estimates of the bid-ask spread available. Moreover, we average monthly spread estimates within the year to further reduce the measurement error. More precisely, $SPREAD_{i,t-1}$ is the average of the monthly spread estimates for stock i in the year (12 months) preceding month t. For clarity of notation, we write the stock-month spread estimate $S_{i,t}$ and the $SPREAD_{i,t-1}$ is given by:

$$SPREAD_{i,t-1} = 1/12 \sum_{n=1}^{12} S_{i,t-n}\,. \tag{2.2}$$

We follow the convention in the asset pricing literature (Cochrane, 2005) that distinguishes between "price of risk", "amount of risk", and "risk premium". However, we adopt a slightly different terminology to make clear that risk and (il)liquidity are two different concepts (Amihud, 2002). We refer to λ as the "price of illiquidity". The stock characteristic *SPREAD* itself captures the "amount of illiquidity". The product λ*SPREAD* is the "illiquidity premium". We notice that some of the earlier contributions on liquidity premia in stock returns use "price" and "premium" interchangeably, effectively normalizing the characteristic to one. However, some contributions explicitly differentiate between the two, in particular Ben-Rephael et al. (2015).

Our selection of control variables ($\mathbf{X}_{i,t-1}$) mainly follows Amihud (2002) and Ben-Rephael et al. (2015). The standard deviation of daily returns (*SDRET*) as well as the logarithm of market equity (*ME*) have been shown to be significant return predictors and are at the same time related to the spread (Stoll, 1978, 1983; Amihud & Mendelson, 1986).[6] The short-term reversal (*STR*) and momentum (*MOM*) have been associated with trading frictions and limits to arbitrage (Kaul & Nimalendran, 1990; Lesmond et al., 2004).

In the robustness tests, we also include control variables based on COMPUSTAT: the logarithm of the book-to-market ratio (*BE/ME*, Fama & French (1992)) and gross profitability (*GP*, Novy-Marx (2013)). Due to the data coverage of COMPUSTAT, these characteristics are only available in the sample period after July 1963. When available, the accounting-based characteristics correlate little with the bid-ask spread and do not alter our results.

2.3 Amount of Illiquidity

We start by illustrating the difference between our measure of bid-ask spreads and popular proxies of transaction costs (*e.g.*, Roll, 1984) or illiquidity more in general (Amihud, 2002).

Figure 2.1 displays the time-series of the *SPREAD* for an equally-weighted portfolio. The *SPREAD* is high in the 1930s, spikes in 1933 (6.27%), it decreases until the 1960s (about 1%), and it increases again with a first peak in 1963 (2.08%), a second peak in 1975 (3.25%), and a third peak in 1992 (5.83%). In line with the idea that liquidity evaporates in times of crisis (Nagel, 2012), these years coincide with periods of financial downturn and economic recession, such as the Great Depression between 1929–41, the U.S. Banking Crisis of 1933, the Kennedy Slide of 1962; the 1973–1975 recession following the oil crisis, the early 1990s recession in the United States, and the 1992 Sterling crisis in Europe. Following the electronization of financial markets in the 2000s, the *SPREAD* has decreased significantly until the global financial crises, when it spikes again in 2009 (1.92%). In the last decade, the *SPREAD* has continued to reduce and it has reached the lowest level ever as of December 2021 (0.36%).

[Insert Figure 2.1 about here.]

The same figure reports bid-ask spreads as measured by the Roll (1984) estimator. This measure overlaps closely with the *SPREAD* in the sample period before 1960, but it underestimates the *SPREAD* between 1960–2000, and overstates it after 2000. The overestimation in the more recent sample is substantial. The estimator fails to capture

[6] See Amihud (2002) for a more detailed discussions and further references.

the downward trend in transaction costs and it produces an estimate of 1.64% as of December 2021, which is about four times larger than the *SPREAD* (Jahan-Parvar & Zikes, 2019; Tremacoldi-Rossi & Irwin, 2019). This problem worsens for larger stocks as displayed in Figure 2.2. The average value-weighted *SPREAD* is 0.05%, while the Roll (1984) measure gives a twentyfold estimate of about 1.00% as of December 2021. In the Appendix (Figures 2.12 and 2.13), we also report bid-ask spreads obtained with the estimators proposed in Corwin & Schultz (2012) and Abdi & Ranaldo (2017). These methods produce estimates closer to the *SPREAD* in the more recent sample period, but they underestimate the *SPREAD* in the past century by a factor of two and are still unable to reproduce the downward trend after 2000. In other words, the bias in bid-ask spreads is not "diversifiable" by using different estimators (*e.g.*, as robustness tests), as they share the common feature of being downward biased where transaction costs are expected to be the largest and upward biased when transaction costs are expected to be the smallest (Ardia et al., 2021).

[Insert Figure 2.2 about here.]

Figures 2.1 and 2.2 also show the evolution of *ILLIQ* for an equally-weighted or value-weighted portfolio, respectively. As the level of *ILLIQ* is not comparable over the years, it is common practice to replace it with its mean-adjusted value, that is, diving *ILLIQ* by its cross-sectional mean (Amihud, 2002). In the next section, we propose an alternative scaling that links *ILLIQ* to the bid-ask spread.

2.3.1 Relation with ILLIQ

In this section, we study the relation between bid-ask spreads and *ILLIQ* (Amihud, 2002). We start by recalling the definition of *ILLIQ* as the ratio of the daily absolute return $|r_{i,t}|$ to the (dollar) trading volume $V_{i,t}$ on that day:

$$ILLIQ_{i,t} = \frac{|r_{i,t}|}{V_{i,t}}\,, \tag{2.3}$$

where i is the stock index and t is the time index.

Giles (2008) shows that the absolute return $|r_{i,t}|$ is an (upward biased) estimator of quadratic volatility $\sigma^2_{i,t}$. This estimator is motivated by the empirical findings of Ghysels et al. (2006) and the theoretical results of Forsberg & Ghysels (2007), that absolute returns outperform squared returns as a predictor of quadratic volatility. Therefore, we write:

$$|r_{i,t}| = k_1 \sigma^2_{i,t}\,, \tag{2.4}$$

where k_1 is a constant coefficient.

Wyart et al. (2008) obtain a linear relation between the volatility per trade and the spread $S_{i,t}$, that must hold universally to reach equilibrium between the profitability of market-making and market-taking strategies:

$$S_{i,t} = k_2 \frac{\sigma_{i,t}}{\sqrt{N_{i,t}}}\,, \tag{2.5}$$

where k_2 is a constant coefficient, and $N_{i,t}$ is the number of trades within day t for stock i.

Substituting Equations (2.4)–(2.5) in (2.3), we obtain:

$$ILLIQ_{i,t} = \frac{k_1}{k_2^2} \frac{S_{i,t}^2}{V_{i,t}/N_{i,t}} \equiv k \frac{S_{i,t}^2}{\mathcal{T}_{i,t}}\,, \tag{2.6}$$

where $k = k_1/k_2^2$ is a coefficient of proportionality, and $\mathcal{T}_{i,t} = V_{i,t}/N_{i,t}$ is the average trade size during day t for stock i (in dollar units). Equation (2.6) shows that *ILLIQ* formally measures the squared bid-ask spread divided by the average trade size in dollar units. As dollar units are not comparable across long time horizons, such as those used for standard asset pricing tests, we express trade sizes in real terms ($\tau_{i,t}$) by adjusting $\mathcal{T}_{i,t}$ for inflation:

$$\tau_{i,t} = \frac{\mathcal{T}_{i,t}}{CPI_t}\,. \tag{2.7}$$

Using Equation (2.7) to rewrite (2.6) in real terms, we obtain:

$$S_{i,t}^2 = \frac{\tau_{i,t}}{k} \times ILLIQ_{i,t} \times CPI_t\,. \tag{2.8}$$

Assuming that the trade size in real terms does not vary widely ($\tau_{i,t} = \tau$), Equation (2.8) states that the bid-ask spread is proportional to the square root of *ILLIQ* adjusted for inflation. We estimate the coefficient of proportionality (τ/k) by computing the median of the ratio $S_{i,t}^2/(ILLIQ_{i,t} \times CPI_t)$ using all the stock-months in CRSP with positive estimates for both the spread and *ILLIQ*. We obtain a coefficient of about 4, so that Equation (2.8) becomes:

$$S_{i,t} = 2\sqrt{ILLIQ_{i,t} \times CPI_t}\,. \tag{2.9}$$

where *ILLIQ* is computed by measuring volume in dollars (and not in millions).[7]

Figure 2.3 displays the comparison between the monthly spread estimates used in this paper and the spreads derived from *ILLIQ* as in Equation (2.9), for an equally weighted portfolio. The two measures overlap with remarkable accuracy and their time-series correlation is above 90%. We also extend the comparison to individual stocks. To this end, we compute the average $ILLIQ_{i,t}$ for each stock i in year t, and we compare $2\sqrt{ILLIQ_{i,t} \times CPI_t}$ with the corresponding $SPREAD_{i,t}$. The overall correlation is 83.5%. This strong link between *ILLIQ* and bid-ask spreads implies that *ILLIQ* can be seen as a non-linear proxy of effective bid-ask spreads and that sorting by *ILLIQ* or by bid-ask spreads is equivalent in the cross-section. Indeed, as CPI_t is fixed in the cross-section, and the square root is a monotonic function, the cross-sectional ranking is theoretically the same whether computed with $ILLIQ_{i,t}$ or with $S_{i,t}$. Empirically, there can be still a difference, particularly when the two are measured with a different degree of error.

[Insert Figure 2.3 about here.]

[7]If volume is measured in million dollars as in Amihud (2002), the coefficient of proportionality in Equation (2.9) becomes 0.002 instead of 2.

2.4 Price of Illiquidity

2.4.1 Baseline Results

Table 2.2 shows results of Fama-MacBeth regressions when we regress monthly excess returns on the *SPREAD*. The first column provides the baseline results with all four control variables. The *SPREAD* has an economically sizeable price of illiquidity of 0.22 and is highly significant (t-statistic: 6.9). Holding all other characteristics equal, a stock with a spread of exactly 1.00% comes with a 0.22% larger monthly return compared to an ideal zero-spread asset. This number implies that investors expect to be compensated for transaction costs within about 4.5 months (1/0.22). This result differ substantially from the estimates obtained by Hasbrouck (2009), who finds that "the gross expected monthly return impounds triple or quadruple the effective cost" and this effect is "too large to be consistent with straightforward trading explanations" as it would imply that investors are compensated for transaction costs within about one week. Our results are more in line with the estimate of 0.21 originally reported in the sample period 1961–1980 by Amihud & Mendelson (1986).

[Insert Table 2.2 about here.]

The table also shows that the regression coefficient for the *SPREAD* is mainly unaffected when removing step-wise the control variables, except for the univariate specification, which does not include volatility as a control variable. Here we find a drop in the coefficient by about one-half. This result is not surprising, as volatile stocks are associated with low average returns and high spreads in the data (Stoll, 1978; Amihud, 2002, 2019). The univariate regressions confounds the two opposing effects.

In the Appendix (Table 2.6), we report results from Fama-MacBeth regressions that include the book-to-market ratio and gross profitability as control variables (Fama & French, 1992; Novy-Marx, 2013). Estimates on the *SPREAD* remain largely unchanged. Since the accounting-based control variables restrict the sample to the period after 06/1963, we do not include them in the baseline results. We also report the results obtained by lagging spreads and volatility by one additional month (Table 2.7) to avoid the short-term reaction of stock returns following unusually large shocks of illiquidity or volatility as discussed by Amihud & Noh (2021b). As our *SPREAD* measure is the average spread in the previous year, we find that lagging by one additional month has essentially no impact and the results are fully consistent.

2.4.2 January Effect

The last two columns of Table 2.2 provide the results when we split the sample in January and all other months. We find that the coefficient on the *SPREAD* is substantially larger in January months (λ: 0.90; t-statistic: 6.3). However, we are able to conclude that the *SPREAD* remains a strong predictor of stock returns also from February to December (λ: 0.15; t-statistic: 5.0).

2.4.3 Trends and Sub-samples

In Figure 2.4, we report the rolling average of λ_t using a 10-year rolling window (including all controls from Table 2.2). The horizontal dashed line corresponds to the full-sample estimate of 0.22 and the shaded area indicates the 90% confidence interval of the rolling sub-sample estimates.

[Insert In Figure 2.4 about here.]

Earlier research documents a significant downward time trend in the price of illiquidity (Ben-Rephael et al., 2015; Drienko et al., 2019; Harris & Amato, 2019). Against this backdrop, we cannot confirm this finding for the *SPREAD*. The figure indicates no important time-trends. In stark contrast to previous research, the 10-year estimates oscillate rather symmetrically around the full sample mean. The 90% confidence intervals usually do not include zero, and almost always include the full-sample estimate of 0.22. In the most recent sub-samples, the point estimate is even above the full sample estimate.

Table 2.3 provides numerical results for the sub-samples 1927-1962, 1963-1999, and 2000-2021. The historical sample (1927-1962) was not examined in the aforementioned literature due to the difficulty in obtaining liquidity measures. The two more recent sub-samples are interesting because the literature commonly finds a positive association between returns and illiquidity measures in the early sample (1963-1999) but not in the recent sample (2000-2021). In line with Figure 2.4, we find that the coefficient for the *SPREAD* is significant and relative stable in all three sub-samples. The table also takes seasonal effects into account. The *SPREAD* remains highly significant in all sub-samples, both in January and non-January months. Only the non-January months in the historical sample come with a relatively low t-statistic of 1.9.

[Insert Table 2.3 about here.]

2.4.4 Size Analysis

In this section, we study how the pricing of bid-ask spreads varies with firm size. To this end, we run the following Fama-MacBeth regression:

$$R_{i,t} - R_{F,t} = a_t + \sum_{m=1}^{M} \lambda_t^{(m)} D_{i,t-1}^{(m)} SPREAD_{i,t-1} + \mathbf{b}_t' \mathbf{X}_{i,t-1} + \varepsilon_{i,t} \, , \tag{2.10}$$

where $D_{i,t-1}^{(m)}$ is a dummy variable that equals 1 if stock i in in the m-th size group at the end of month $t-1$, and $\lambda_t^{(m)}$ is the corresponding price of illiquidity. We use M size groups partitioned by market capitalization at the end of last June (smaller to larger).

In our first test, we use $M = 3$ to split the sample in small, medium, and large stocks. We find that the price of illiquidity $\lambda^{(m)}$ is statistically significant for all stocks. We obtain a coefficient $\lambda^{(1)} = 0.22$ (t-statistic: 6.95) for small stocks, $\lambda^{(2)} = 0.12$ (t-statistic: 2.71) for medium stocks, and $\lambda^{(3)} = 0.23$ (t-statistic: 2.61) for large stocks.

This result is in constrast with Ben-Rephael et al. (2015), who report a significant liquidity premium only among the smallest size tercile and conclude that the liquidity premium (p.214) “is narrowly concentrated among small stocks.” We cannot share this conclusion for the *SPREAD*.

We repeat the analysis by increasing the number of groups to $M = 25$. Figure 2.5 shows the time-series average and the 95% confidence intervals of the cross-sectional estimates $\lambda_t^{(m)}$ for each size group. We notice that the estimates tend to decline and come with larger dispersion the more we move from small-size groups to large-size groups. We conjecture that this effect is due to statistical reasons. First, bid-ask spreads of larger stocks are typically smaller than those of smaller stocks (Roll, 1984) and their estimation is upward biased due to the practice of resetting negative spread estimates to zero (see Jahan-Parvar & Zikes, 2019; Tremacoldi-Rossi & Irwin, 2019; Ardia et al., 2021). Consequently, the associated regression coefficient is downward biased. Second, the standard deviation of the *SPREAD* in the largest-size group (0.27%) is more limited than that in the smallest-size group (6.16%). As higher variation in the independent variables typically leads to greater precision of the regression coefficients, this implies that the price of illiquidity is estimated more precisely for smaller stocks.

[Insert Figure 2.5 about here.]

To verify our conjecture, we run a simulation experiment. For each stock i and for each month t, we simulate a monthly return equal to $R_{i,t} = 0.22 \times SPREAD_{i,t} + \varepsilon_{i,t}$, where $SPREAD_{i,t}$ is the empirical spread estimate and $\varepsilon_{i,t}$ is distributed normally with mean zero and standard deviation equal to the empirical monthly volatility. The monthly volatility is computed by multiplying the standard deviation of the empirical daily returns within the month times the square root of 21 (average number of trading days per month). We split stocks in size groups and we estimate the regression in Equation (2.10) using the simulated returns (and without control variables). The spread used in the regression is $\max\{0, SPREAD_{i,t} + u_{i,t}\}$, where $u_{i,t}$ is a measurement error distributed normally with mean zero and standard deviation σ_u. Figure 2.6 reports the results. When the measurement error is zero, the estimates for larger stocks come with larger estimation uncertainty and do not decline. When the measurement error increases, the estimates of the larger stocks shrink to zero.

[Insert Figure 2.6 about here.]

In summary, our results suggest that the *SPREAD* is priced across all stocks, although samples restricted to larger stocks lack statistical power to detect the relationship between the *SPREAD* and expected returns. For this reason, it becomes particularly important to use an accurate measure of the bid-ask spread. Our results show that the spread is priced among all size terciles, including the 33% of the largest stocks.

2.5 Illiquidity Premium

2.5.1 The Premium in Percentage Points

In this section, we study the evolution of the illiquidity premium measured in percentage points per month:

$$\Pi_{\%,t} = \sum_i \hat{\lambda}_t SPREAD_{i,t-1} \times w_{i,t-1}, \tag{2.11}$$

where $\hat{\lambda}_t$ is the estimate of the price of illiquidity in month t using the specification in Equation (2.1), and $w_{i,t}$ are the weights of the individual stocks summing up to one. We use equal and value weighting to measure the average premium and the market portfolio's premium, respectively.

Figure 2.7 shows the 10-year rolling average of $\Pi_{\%,t}$ and the corresponding 90% confidence intervals. Over the full sample (indicated by dashed lines), the average premium (equally weighted) is 0.32% per month. For the market portfolio (value weighted), we estimate a premium of 0.09% per month. The lower premium for the market reflects the fact that stocks with a large market capitalization tend to have smaller spreads (Roll, 1984). The premium of the market portfolio in annual terms is 1.1%. The equity premium is usually estimated to be around 6% (Kroencke, 2017). Accordingly, we can attribute an economically significant share of about 20% of the historical equity premium to compensation for liquidity.

[Insert Figure 2.7 about here.]

We find that the premium for the market portfolio has declined over the last decade. The most recent estimate of the premium at the end of the sample is still positive and marginally significant, but it is lower than the full sample mean. Because the price of illiquidity λ is stable over time (see Figure 2.4), or even somewhat increasing, the decline must come from decreasing *SPREAD* (see Figure 2.2) or an increase in the weights of the market portfolio towards stocks with a lower *SPREAD*. In both cases, our results indicate that the value-weighted "amount of illiquidity" has changed, while the "price of illiquidity" has remained stable.

Interestingly, we do not find a clear time-trend for the average premium. At the end of our sample period, the premium for the equally weighted portfolio is close to the full sample mean and highly significant. This is in contrast with the idea that liquidity premia have generally declined (Ben-Rephael et al., 2015).

2.5.2 The Premium in Real Dollars

While the premium in percentage points provides a measure of the marginal costs of liquidity, it says nothing about the total compensation required by investors in the equity market. For this reason, we study the premium measured in real dollars per month:

$$\Pi_{\$,t} = \sum_i \hat{\lambda}_t SPREAD_{i,t-1} \times ME_{i,t-1} \frac{CPI_0}{CPI_{t-1}} \times w_{i,t-1}\,, \tag{2.12}$$

where $\hat{\lambda}_t$ is the estimate of the price of illiquidity in month t using the specification in Equation (2.1). $ME_{i,t-1}$ is the market capitalization of stock i in month $t-1$, and it is adjusted by inflation using the consumer price index (CPI). We set CPI_0 equal to the value as of December 2021 to express dollar units in real terms as of the end of our sample. Finally, $w_{i,t}$ are the weights of the individual stocks summing up to one. We use equal and value weighting. In words, $\Pi_{\$,t}$ is the real dollar amount that investors require per stock per month to compensate for transaction costs as measured by the *SPREAD*.

Figure 2.8 reports the 10-year rolling average of $\Pi_{\$,t}$ and the corresponding 90% confidence intervals. The dashed lines indicate the full sample averages. The average equally-weighted premium is 1.5 million per stock-month and it increases to 6.8 millions for the value-weighted average. This result illustrates that the lower *SPREAD* of large-cap stocks actually adds up to a large dollar amount when moving from percentage returns to total compensation in dollars.

[Insert Figure 2.8 about here.]

Furthermore, Figure 2.8 illustrates that the premium in real dollars has no clear time-trend for the equally-weighted average as well as for the value-weighted average. An increasing real market capitalization counters the decrease of the *SPREAD* of the larger stocks over time. This finding could be coincidence. But it is also in line with the hypothesis that investors demand a constant absolute compensation for transaction costs in contrast to a constant relative compensation.

To assess the total effect on the market, we compute $\Pi_{\$,t}$ in Equation (2.12) with $w = 1$. That is, we sum the dollar premium for all stocks. Figure 2.9 reports the results. The top panel reports the time-series of the dollar premium for the whole market. The bottom panel shows the average number of stocks per month. We find that the total premium is statistically significant and economically sizeable throughout the sample. It increases from about 0.65 billion dollars per month in 1936 (600 stocks) to 15 billion dollars around 2000 (6000 stocks). In the last decade, we estimate a premium of 5.67 billions per month (about 3500 stocks). In the full sample, we find an average premium of 4.12 billion dollars per month.

[Insert Figure 2.9 about here.]

2.6 Clientele Effects

In this section, we study how the price of illiquidity and the illiquidity premium vary across spread sizes. The seminal work by Amihud & Mendelson (1986) presents a model where investor demand compensation for holding assets that are more expensive to trade and, as a consequence the asset's return increases in the bid-ask spread. The model leads to two main predictions. First, assets with higher spreads are allocated to investors with longer holding periods (clientele effect). Second, the gross return is an increasing and concave function of the spread (spread-return relationship).

According to the model, the gross return required by investor j on asset i is given by:

$$R_{i,j} = r_j^* + \mu_j S_i \,, \tag{2.13}$$

where $1/\mu_j = \mathbb{E}[T_j]$ is the expected holding period of investor j, S_i is the bid-ask spread of asset i, and r_j^* is the required net return (Amihud & Mendelson, 1986, Equation 5).

We test the model's predictions using the following specification:

$$R_{i,t} - R_{F,t} = a_t + \sum_{n=1}^{10} \lambda_t^{(n)} D_{i,t-1}^{(n)} SPREAD_{i,t-1} + \mathbf{b}'\mathbf{X}_{i,t-1} + \varepsilon_{i,t} \tag{2.14}$$

where $D_{i,t-1}^{(n)}$ is a dummy variable that equals 1 if stock i in in the n-th spread decile at the end of month $t-1$, and $\lambda_t^{(n)}$ is the corresponding price of illiquidity. Equation (2.13) implies that $\lambda_n = \mu_n = 1/\mathbb{E}[T_n]$. That is, the price of illiquidity of the n-th spread decile corresponds to the inverse of the expected holding period of an investor holding the decile. The clientele effect implies that $\lambda^{(n)}$ should be positive and it should decrease as we move to higher-spread deciles.

Figure 2.10 reports the estimates of $\lambda^{(n)}$ for each spread decile. All estimates are positive and statistically significant, confirming the idea that the spread is priced for each level of the spread. In agreement with the clientele effect, the coefficients $\lambda^{(n)}$ declines as we move to higher-spread groups. In the context of the Amihud & Mendelson (1986)-model, we find that the smallest-spread stocks are associated with a holding period of about one week ($4 < \lambda^{(1)} < 5$), while the highest-spread stocks are associated with a holding period of about 4 months ($\lambda^{(10)} \approx 0.25$). Average-spread stocks have a holding period of about one month ($\lambda^{(5)} \approx 1$).

[Insert Figure 2.10 about here.]

We now turn to the second prediction: the gross return is an increasing and concave function of the spread.[8] Intuitively, less liquid assets are allocated to investors with longer holding periods (smaller λ), which mitigates the compensation that they require for the costs of illiquidity (Amihud et al., 2006). We test this prediction as follows. Let $\lambda_t^{(n)}$ be the coefficient of the n-th decile in Equation (2.14) for the cross-sectional regression in month t. For each month, we compute the premium $\Pi_{\%,t}^{(n)} = \lambda_t^{(n)} SPREAD_{t-1}^{(n)}$, where $SPREAD_{t-1}^{(n)}$ denotes the average spread in the n-th decile. According to the spread-return relationship, $\Pi_{\%,t}^{(n)}$ should be concave and increasing from lower to higher spread deciles. Finally, we compute $\Pi_{\%}^{(n)}$ as the average of $\Pi_{\%,t}^{(n)}$ across months. As monotonicity and concavity are preserved by the sum operator, $\Pi_{\%}^{(n)}$ should be monotonically increasing and concave.

Figure 2.11 reports the estimates of $\Pi_{\%}^{(n)}$ for each spread decile. We find that the premium is generally increasing with the spread, but it is not concave. The estimates are flat from decile 1 to deciles 8 (about 0.30%), start increasing for decile 9 (about 0.50%), and are significantly higher for the highest-spread stocks (about 1.25%), giving raise to a convex spread-return relationship. We notice that the average spread in decile 10 (8.03%) is, on average, 2–3 times larger than deciles 8–9 (2.71%–4.05%), while their expected holding periods do not differ substantially (see Figure 2.10). In other words, the higher premium required by the highest-spread stocks does not seem to be explained in terms of compensation for transaction costs alone.

[Insert Figure 2.11 about here.]

To further test the spread-return relationship, we regress the monthly excess returns on the squared spread ($SPREAD^2$) and on the square root of the spread ($\sqrt{SPREAD}$). The estimates and their t-statistics are given below. We find that only the coefficient

[8] The case of a flat spread-return relationship is allowed.

on $SPREAD^2$ is significant, while $\sqrt{SPREAD}$ is not, and the regression selects a convex relationship.

$$R_{i,t} - R_{F,t} = a + \underset{[4.46]}{361} SPREAD^2_{i,t-1} - \underset{[0.64]}{0.59} \sqrt{SPREAD_{i,t-1}} + \mathbf{b}'\mathbf{X}_{i,t-1} + \varepsilon_{i,t}\,. \tag{2.15}$$

In summary, while our results are in line with a concave (flat) spread-return relationship for 90% of the stocks, the premium of the top decile is too high to be explained solely in terms of compensation from transaction costs.

2.7 Traded Factor

In Section 2.3.1, we have shown that *ILLIQ* (Amihud, 2002) is closely related the bid-ask spread and that cross-sectional rankings based on *ILLIQ* are equivalent to sorting on the spread. Here we replicate the "Illiquid minus Liquid" (IML) factor (Amihud et al., 2015; Amihud, 2019; Amihud & Noh, 2021b) by replacing portfolio sorts on *ILLIQ* with sorts on the *SPREAD*. We call this factor "Expensive minus Cheap" (EMC). While IML captures liquidity in general and has no unique economic interpretation (see *e.g.*, Amihud, 2002; Chordia et al., 2009), EMC is directly linked to transaction costs: it goes long in stocks that are expensive to trade (*i.e.*, large bid-ask spreads) and goes short in stocks that are cheap to trade (*i.e.*, small bid-ask spread).

To construct the IML factor we proceed as follows. At the beginning of the month, we sort stocks into three portfolios by their volatility (*SDRET*) in the previous month. Then, within each volatility tercile, stocks are sorted into quintiles by their *ILLIQ* measure. To calculate returns, we use both value and equally weighted portfolios. The IML factor is the return of the illiquid-minus-liquid quintile portfolios, averaged across the volatility portfolios (see Amihud et al., 2015). EMC replaces the *ILLIQ* measure with the *SPREAD* and follows the same construction methodology.

Table 2.4 reports the comparison between IML and EMC. The results are similar and practically equivalent. In the full sample, the average value-weighted monthly return for EMC (IML) is 0.34% (0.30%) with a t-statistic of 2.72 (2.46). The standard deviation of monthly returns is 4.23% (4.08%). The alpha with respect to the Fama-French three factor model is 0.25% (0.34%) with a t-statistic of 2.65 (3.80).[9] Also the loadings on the market factor (β_{MKT}), small minus big (β_{SMB}), and high minus low (β_{HML}), are similar. We obtain -0.32 (-0.45) for β_{MKT}, 0.69 (0.61) for β_{SMB}, and 0.50 (0.42) for β_{HML}. The results are also similar when the portfolios are equally weighted. In this case, the average return increases to 0.77% (0.75%) per month, while the standard deviation of returns stays stable at 4.28% (4.43%). The table further reports results for the historical sample 1927–1962, the early sample 1963–1999, and the more recent sample 2000–2021. IML and EMC produce similar results in all sub-periods, in terms of summary statistics, significance levels, alpha, and factor loadings, both for equal and value weighted portfolios.

[Insert Table 2.4 about here.]

[9] We use the three factor model from the website of Kenneth R. French.

Table 2.5, Panel A, regresses IML on EMC. In the full sample, we find that α_{EMC} is small and statistically insignificant for both the value-weighted and the equally-weighted factors, meaning that IML does not yield additional returns compared to EMC. β_{EMC} is highly significant and the R^2 is up to 84%, implying that the EMC explains most of the variation of IML. The results are similar across all sub-samples. Only in the equally weighted regression for the period 1963–1999 IML performs marginally worse than EMC ($\hat{\alpha}_{\text{EMC}} = -0.16$; t-statistic: -2.37). Panel B regresses EMC on IML and yields the same conclusions.

[Insert Table 2.5 about here.]

In summary, our results show that EMC closely replicates IML. This supports the idea that cross-sectional sorts based on *ILLIQ* are equivalent to cross-sectional sorts based on bid-ask spreads, as predicted by Equation (2.9). Thus, we find that the premium of the IML factor can be explained in terms of compensation from transaction costs (Amihud & Mendelson, 1986).[10] This interpretation may apply more generally to applications that rely on cross-sectional sorts on the Amihud (2002) illiquidity measure. However, we highlight that bid-ask spreads and *ILLIQ* are two different measures and we do not expect them to produce similar results in other types of analyses. Equation (2.9) shows that the bid-ask spread is a non-linear transformation of *ILLIQ*, it incorporates inflation, and the coefficient of proportionality may generally depend on the market and asset class. While cross-sectional rankings based on the two measures are expected to be similar, there may be significant differences for time-series studies, analyses of international markets, regression studies, and when the magnitude of the liquidity measure matters (see *e.g.*, Acharya & Pedersen, 2005; Ben-Rephael et al., 2015; Novy-Marx & Velikov, 2016).

2.8 Conclusion

This paper studies liquidity premia in the U.S. stock market since 1927. The liquidity measure used in this paper is the effective bid-ask spread to capture the effective cost of trading as in the seminal model by Amihud & Mendelson (1986). We find that: (a) liquidity is more strongly priced in January, but it is also significant in non-January months; (b) liquidity is priced across all stocks, although samples restricted to larger stocks may lack power and achieve lower significance; (c) stocks with higher spreads are associated with lower prices of illiquidity and higher illiquidity premia.

Our results differ from previous research that suggests that liquidity is priced only in January and liquidity premia are narrowly concentrated among small stocks. Instead, we find that liquidity is significantly priced throughout the year and across stocks. Moreover, we find that liquidity premia are still large and significant in the recent years, in contrast with the idea that they have generally declined over time and have become a second-order effect.

This work provides strong empirical support for the predictions in Amihud & Mendelson (1986). Specifically, our results are in line with the idea that assets with higher spreads are allocated to investors with longer holding periods and that the premium is

[10]See discussion in Lou & Shu (2017); Amihud & Noh (2021a).

an increasing function of the spread. We also show that the popular Amihud (2002) illiquidity measure can be seen as a non-linear proxy for bid-ask spreads.

In summary, we find that bid-ask spreads represent an important dimension of liquidity and they have pervasive effects on asset prices. The heterogeneity of results reported in literature arise from different liquidity measures, which are easier to compute but harder to justify from the perspective of theoretical models. The results we obtain by measuring liquidity with effective bid-ask spreads (Amihud & Mendelson, 1986) are more consistent over time and across stocks, and demonstrate the benefits of adopting liquidity measures motivated by theory and with clear economic interpretation rather than proxies motivated by intuition and that are difficult to interpret (Chordia et al., 2009).

Bibliography

Abdi, F., & Ranaldo, A. (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies*, *30*(12), 4437–4480.

Acharya, V. V., & Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of Financial Economics*, *77*, 375–410.

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

Amihud, Y. (2019). Illiquidity and stock returns: A revisit. *Critical Finance Review*, *8*, 203–221.

Amihud, Y., Hameed, A., Kang, W., & Zhang, H. (2015). The illiquidity premium: International evidence. *Journal of Financial Economics*, *117*(2), 350–368.

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

Amihud, Y., Mendelson, H., Pedersen, L. H., et al. (2006). Liquidity and asset prices. *Foundations and Trends® in Finance*, *1*(4), 269–364.

Amihud, Y., & Noh, J. (2021a). Illiquidity and stock returns II: Cross-section and time-series effects. *Review of Financial Studies*, *34*(4), 2101–2123.

Amihud, Y., & Noh, J. (2021b). The pricing of the illiquidity factor's conditional risk with time-varying premium. *Journal of Financial Markets*, *56*, 100605.

Ardia, D., Guidotti, E., & Kroencke, T. A. (2021). Efficient estimation of bid-ask spreads from open, high, low, and close prices. Working paper, SSRN 3892335.

Ben-Rephael, A., Kadan, O., & Wohl, A. (2015). The diminishing liquidity premium. *Journal of Financial and Quantitative Analysis*, *50*(1/2), 197–229.

Brennan, M., Huh, S.-W., & Subrahmanyam, A. (2013). An analysis of the amihud illiquidity premium. *Review of Asset Pricing Studies*, *3*(1), 133–176.

Chelley-Steeley, P. L., Lambertides, N., & Steeley, J. M. (2015). The effects of non-trading on the illiquidity ratio. *Journal of Empirical Finance*, *34*, 204–228.

Chordia, T., Huh, S.-W., & Subrahmanyam, A. (2009). Theory-based illiquidity and asset pricing. *Review of Financial Studies*, *22*(9), 3629–3668.

Cochrane, J. H. (2005). *Asset pricing*. Princeton [u.a.]: Princeton Univ. Press.

Corwin, S. A., & Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance*, *67*(2), 719–760.

Drienko, J., Smith, T., & von Reibnitz, A. (2019). A review of the return–illiquidity relationship. *Critical Finance Review*, *8*, 127–171.

Eleswarapu, V., & Reinganum, M. R. (1993). The seasonal behavior of the liquidity premium in asset pricing. *Journal of Financial Economics*, *34*(3), 373–386.

Eleswarapu, V. R. (1997). Cost of transacting and expected returns in the nasdaq market. *Journal of Finance*, *52*(5), 2113–2127.

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

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

Forsberg, L., & Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, *5*(1), 31–67.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, *131*(1-2), 59–95.

Giles, D. E. (2008). Some properties of absolute returns as a proxy for volatility. *Applied Financial Economics Letters*, *4*(5), 347–350.

Goyenko, R. Y., Holden, C. W., & Trzcinka, C. A. (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics*, *92*, 153–181.

Harris, L., & Amato, A. (2019). Illiquidity and stock returns: Cross-section and time-series effects: A replication. *Critical Finance Review*, *8*, 173–202.

Hasbrouck, J. (2009). Trading costs and returns for us equities: Estimating effective costs from daily data. *Journal of Finance*, *64*(3), 1445–1477.

Holden, C. W., & Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance*, *69*(4), 1747–1785.

Jahan-Parvar, M. R., & Zikes, F. (2019). When do low-frequency measures really measure transaction costs? Working paper, SSRN 3418948.

Jegadeesh, N., Noh, J., Pukthuanthong, K., Roll, R., & Wang, J. (2019). Empirical tests of asset pricing models with individual assets: Resolving the errors-in-variables bias in risk premium estimation. *Journal of Financial Economics*, *133*(2), 273–298.

Kaul, G., & Nimalendran, M. (1990). Price reversals: Bid-ask errors or market overreaction? *Journal of Financial Economics*, *28*(1), 67–93.

Kroencke, T. A. (2017). Asset pricing without garbage. *Journal of Finance*, *72*(1), 47–98.

Lesmond, D. A., Ogden, J. P., & Trzcinka, C. A. (1999). A new estimate of transaction costs. *Review of Financial Studies*, *12*(5), 1113–1141.

Lesmond, D. A., Schill, M. J., & Zhou, C. (2004). The illusory nature of momentum profits. *Journal of Financial Economics*, *71*(2), 349–380.

Lou, X., & Shu, T. (2017). Price impact or trading volume: Why is the amihud (2002) measure priced? *Review of Financial Studies*, *30*(12), 4481–4520.

Nagel, S. (2012). Evaporating liquidity. *Review of Financial Studies*, *25*(7), 2005–2039.

Novy-Marx, R. (2013). The other side of value: The gross profitability premium. *Journal of Financial Economics*, *108*(1), 1–28.

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *Review of Financial Studies*, *29*, 104–147.

Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, *39*(4), 1127–1139.

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

Stoll, H. R. (1978). The pricing of security dealer services: An empirical study of nasdaq stocks. *Journal of Finance*, *33*(4), 1153–1172.

Stoll, H. R. (1983). Transaction costs and the small firm effect. *Journal of Financial Economics*, *12*(1), 57–79.

Tremacoldi-Rossi, P., & Irwin, S. H. (2019). A problem of zeros: The misbehavior of simple bid-ask spread estimators. Working Paper.

Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57.

Figures

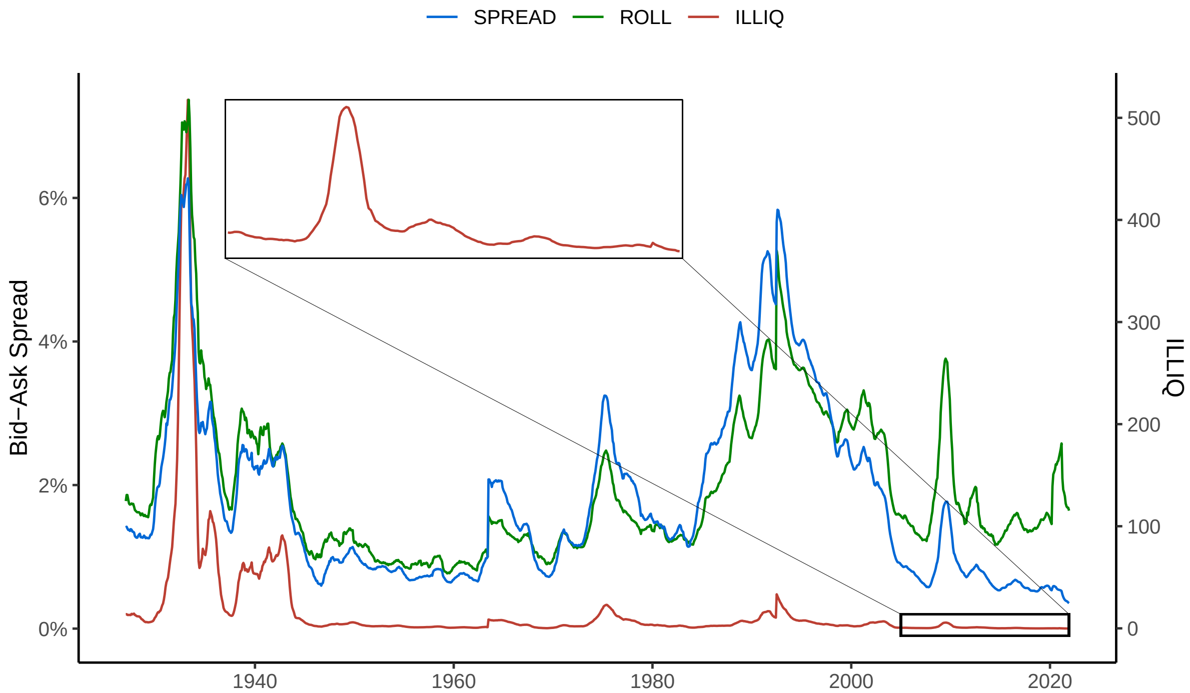


Figure 2.1: Time-series of illiquidity measures for an equally-weighted portfolio. *SPREAD* is the bid-ask spread used in this paper, *ROLL* is the bid-ask spread estimator in Roll (1984), *ILLIQ* is the Amihud (2002) illiquidity measure. Each measure is computed monthly for each stock. Then, it is averaged using a 12-months rolling window for each stock. Finally, it is averaged (equally-weighted) across stocks. The left axis reports the scale for the *SPREAD* and the *ROLL* measures. The right axis reports the scale for the *ILLIQ* measure.

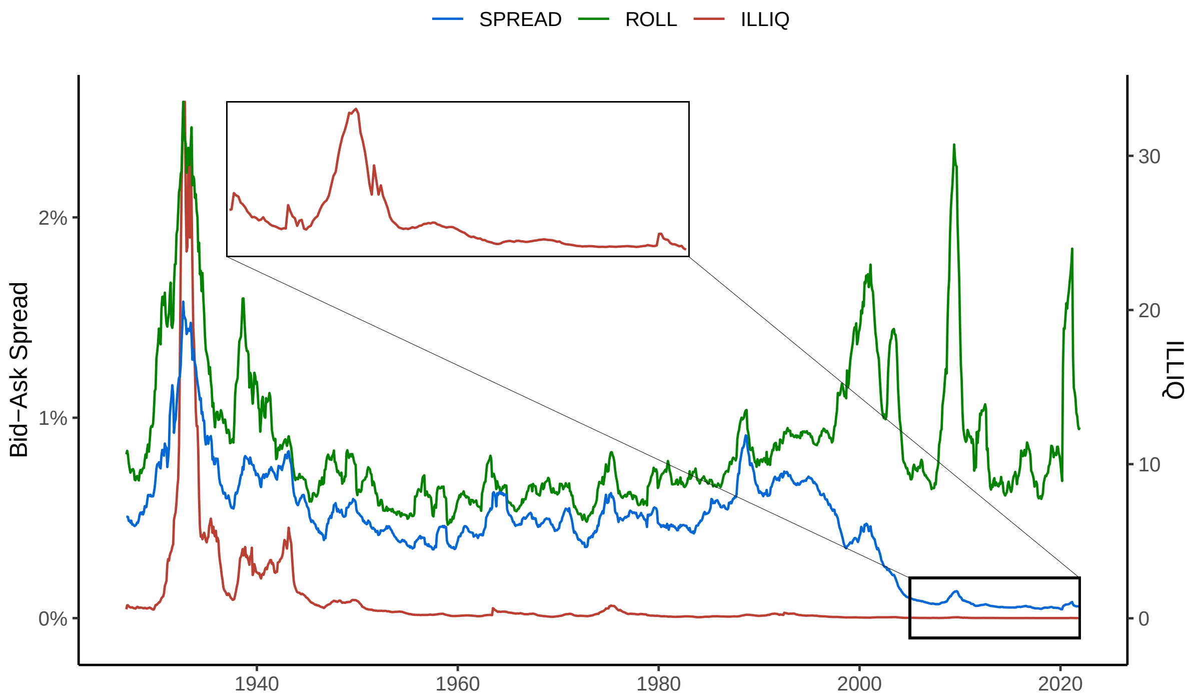


Figure 2.2: Time-series of illiquidity measures for a value-weighted portfolio. *SPREAD* is the bid-ask spread used in this paper, *ROLL* is the bid-ask spread estimator in Roll (1984), *ILLIQ* is the Amihud (2002) illiquidity measure. Each measure is computed monthly for each stock. Then, it is averaged using a 12-months rolling window for each stock. Finally, it is averaged (value-weighted) across stocks. The left axis reports the scale for the *SPREAD* and the *ROLL* measures. The right axis reports the scale for the *ILLIQ* measure.

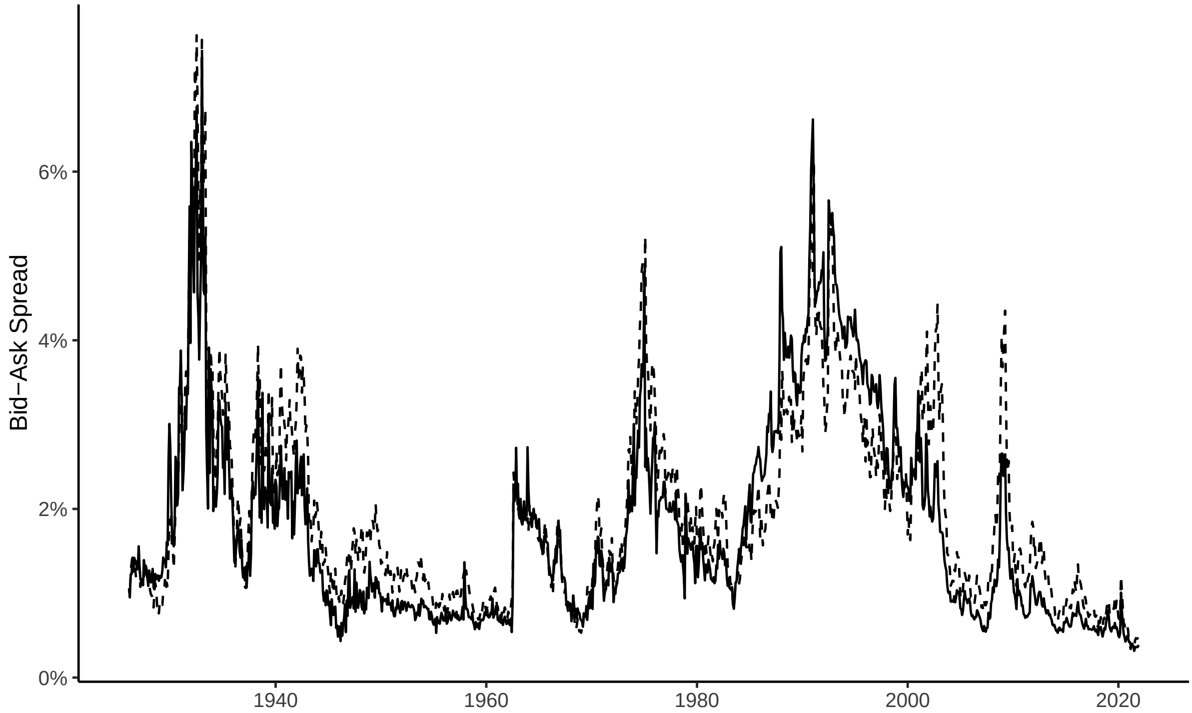


Figure 2.3: Relation with *ILLIQ*. The figure shows the time-series of the average monthly spread estimate. Spreads are computed for each stock-month and averaged across stocks for each month. The solid line corresponds to the bid-ask spreads estimated with EDGE (Ardia et al., 2021) or computed directly from TAQ when intraday data are available. The dotted line represents the spreads computed from *ILLIQ* as in Equation (2.9).

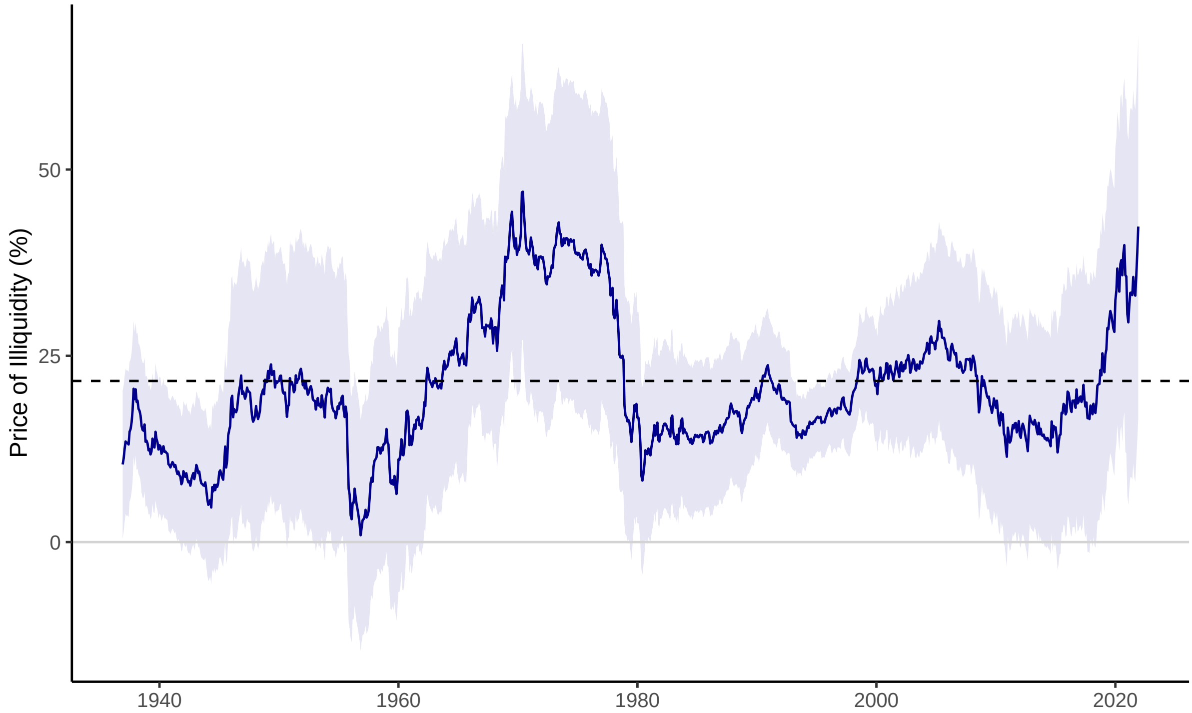


Figure 2.4: Price of illiquidity. The figure shows the 10-year rolling average of the coefficient for the Fama-MacBeth regression number 1 in Table 2.2 (including all controls). The shaded area represents the 90% confidence intervals. The dashed line is the full sample average.

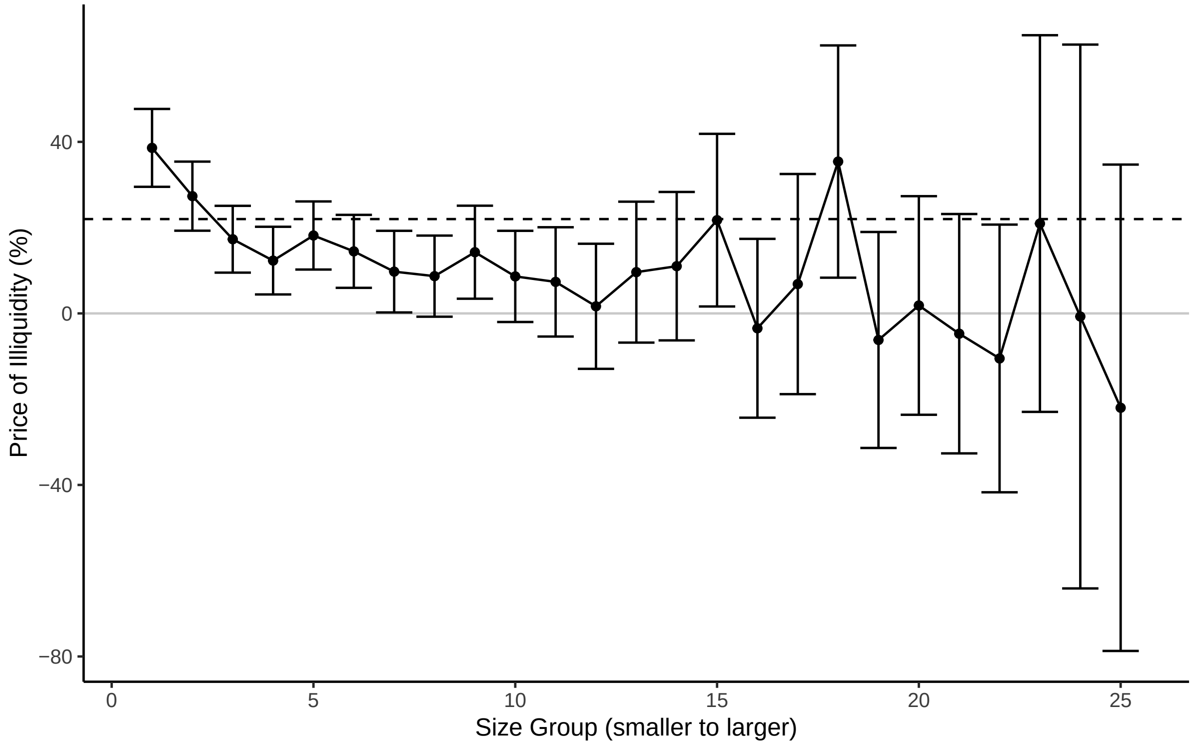


Figure 2.5: Price of illiquidity by size group. The figure shows the time-series average and the 95% confidence intervals of the cross-sectional estimates of $\lambda_t^{(m)}$ in Equation (2.10) for each size group m. The dashed line is the full-sample estimate of 22 percentage points (Table 2.2).

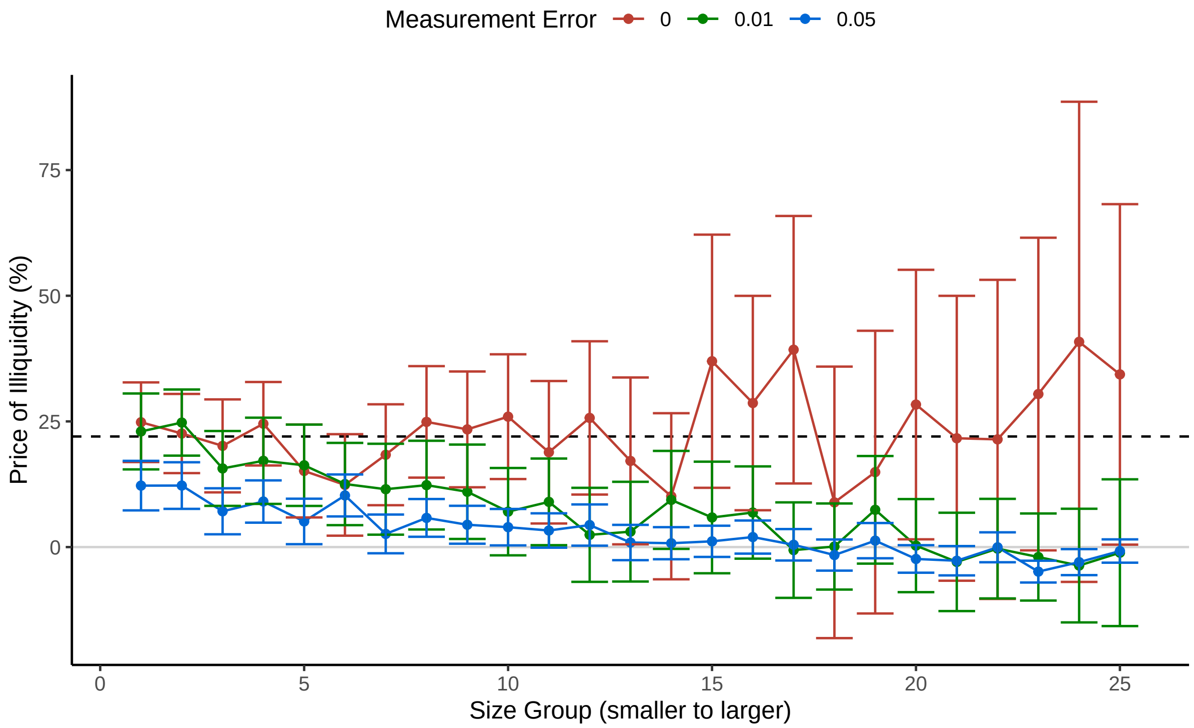


Figure 2.6: Price of illiquidity by size group and measurement error. For each stock i and for each month t, we simulate a monthly return equal to $R_{i,t} = 0.22 \times SPREAD_{i,t} + \varepsilon_{i,t}$, where $SPREAD_{i,t}$ is the empirical spread estimate and $\varepsilon_{i,t}$ is distributed normally with mean zero and standard deviation equal to the empirical monthly volatility. The monthly volatility is computed by multiplying the standard deviation of the empirical daily returns within the month times the square root of 21 (average number of trading days per month). We split stocks in size groups as discusses in Section 2.4.4 and we estimate the regression in Equation (2.10). The spread used in the regression is $\max\{0, SPREAD_{i,t} + u_{i,t}\}$, where $u_{i,t}$ is a measurement error distributed normally with mean zero and standard deviation σ_u. The figure shows the time-series average and the 95% confidence intervals of the cross-sectional estimates $\lambda_t^{(m)}$ in Equation (2.10) for each size group m, obtained for increasing values of σ_u. The dashed line is the price of illiquidity of 22 percentage points used to simulate returns.

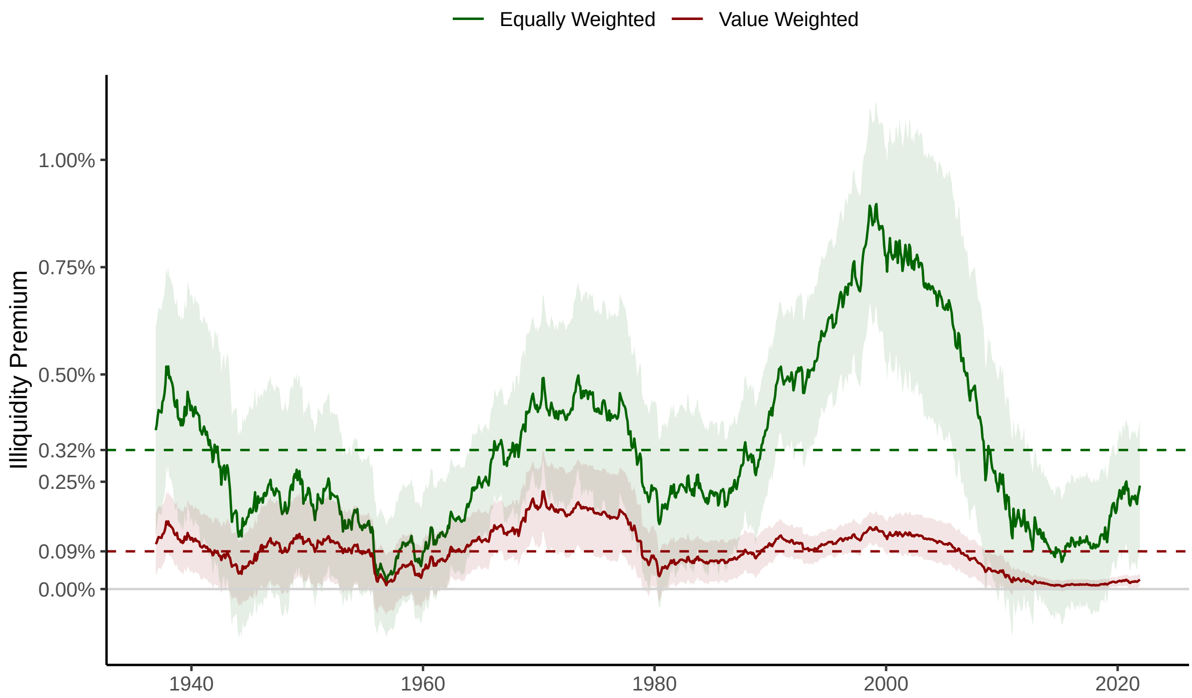


Figure 2.7: Illiquidity premium in percentage points per month. The figure shows the 10-year rolling average of $\Pi_{\%,t}$ in Equation (2.11) and the corresponding 90% confidence intervals for a value-weighted and an equally-weighted portfolio. The dashed lines are the corresponding average premia in the full sample.

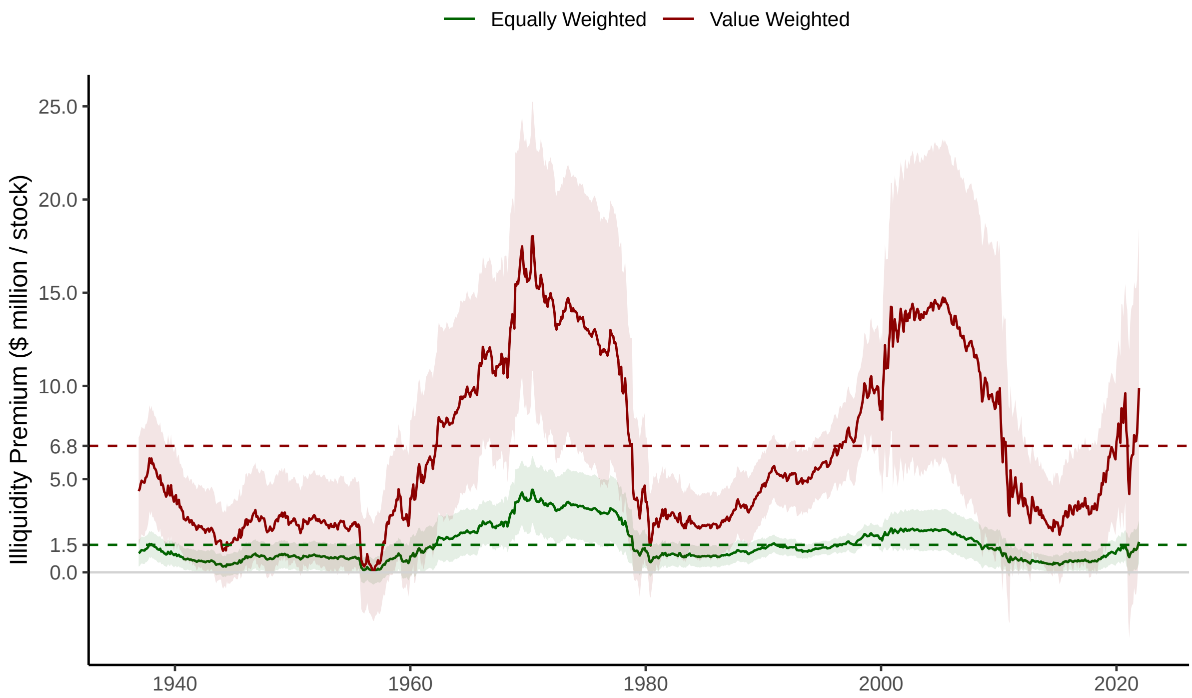


Figure 2.8: Illiquidity premium in (real) million dollars per stock-month. The figure shows the 10-year rolling average of $\Pi_{\$,t}$ in Equation (2.12) and the corresponding 90% confidence intervals. $\Pi_{\$,t}$ is computed using equal weights for each stock (equally-weighted) or weights proportional to the stock's capitalization (value-weighted). The dashed lines are the corresponding average premia in the full sample.

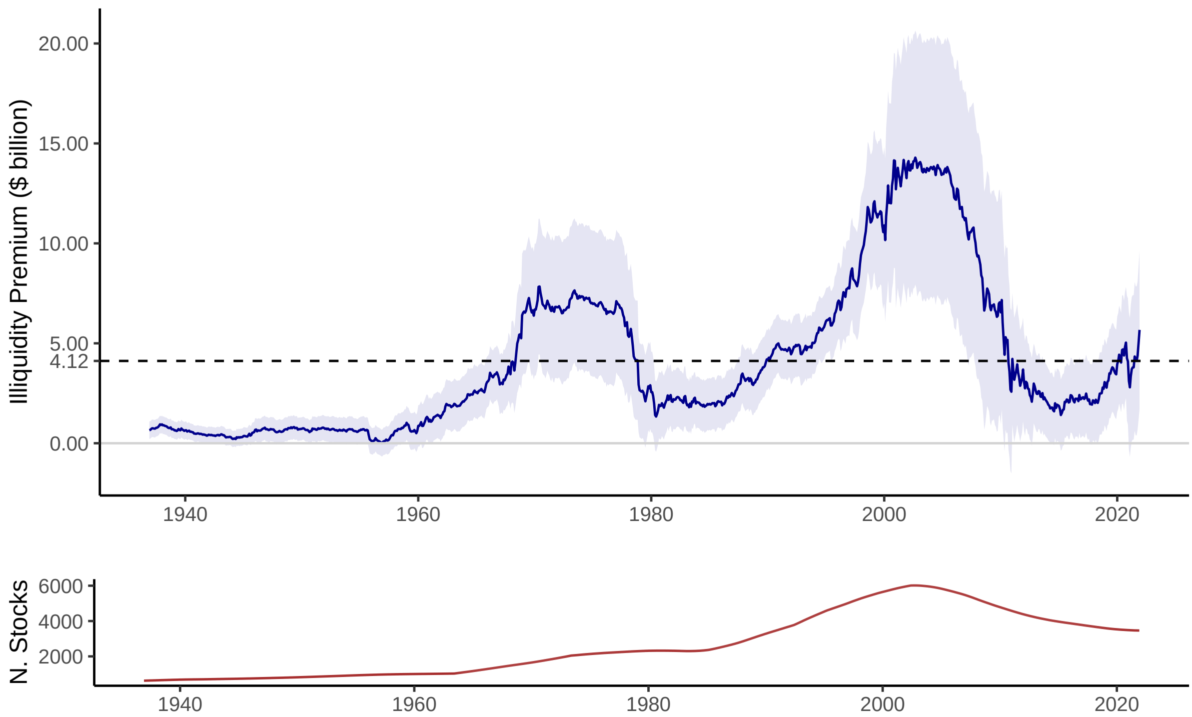


Figure 2.9: Illiquidity premium in (real) billion dollars per month. The top panel shows the 10-year rolling average of $\Pi_{\$,t}$ in Equation (2.12) and the corresponding 90% confidence intervals. The dashed line is the average premium in the full sample. $\Pi_{\$,t}$ is computed using $w = 1$ to sum the dollar premium for all stocks. The bottom panel shows the 10-year rolling average of the number of stocks in the sample (NYSE, AMEX, NASDAQ).

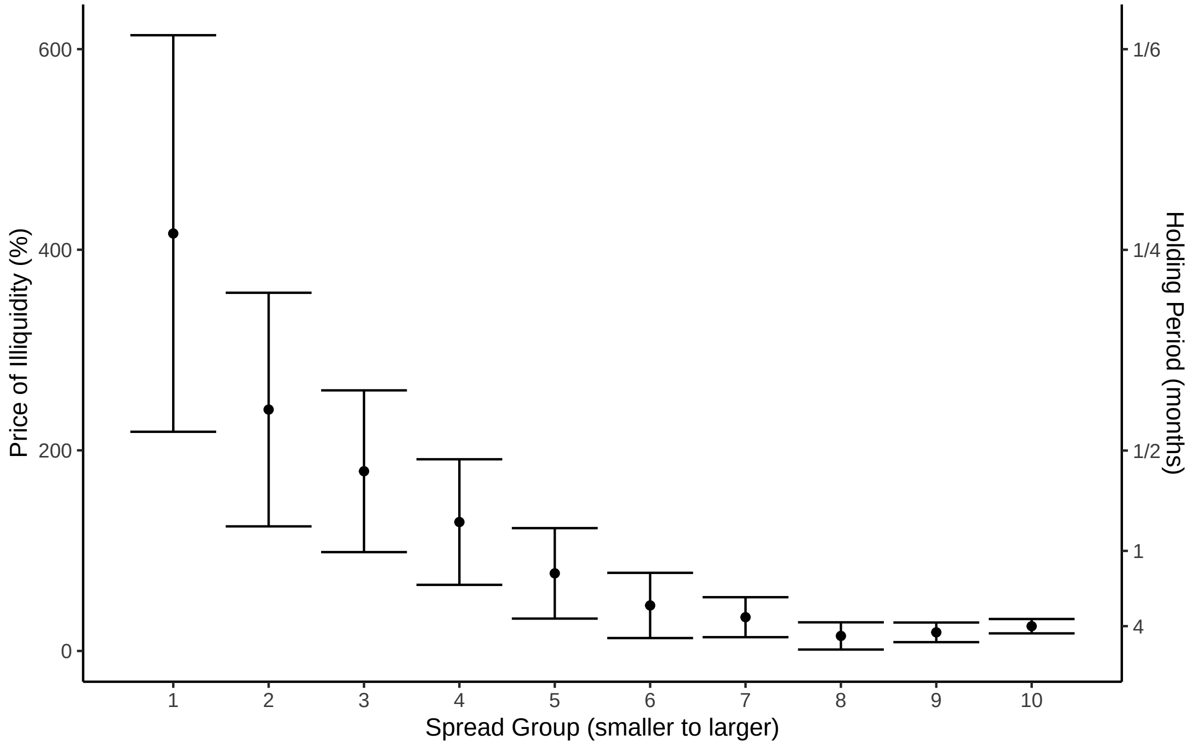


Figure 2.10: Price of illiquidity by spread group. The figure shows the estimates of $\lambda^{(n)}$ in Equation (2.14) for each spread group n. The lower and upper bars indicate the 95% confidence intervals. The right axis reports the holding period corresponding to the estimates.

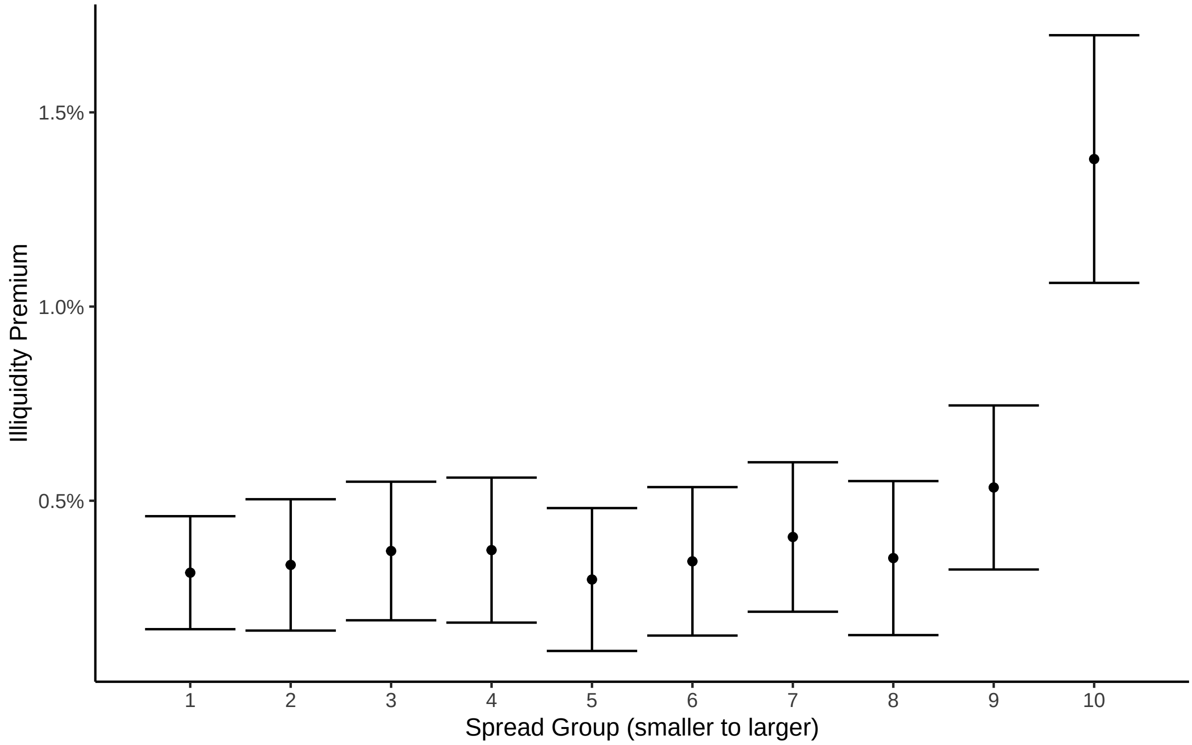


Figure 2.11: Illiquidity premium by spread group. The figure shows the estimates of $\Pi_{\%}^{(n)}$ in Section 2.6 for each spread group n. The lower and upper bars indicate the 95% confidence intervals.

Tables

Variable	N	Mean	Sd	Pctile[10]	Pctile[50]	Pctile[90]
SPREAD (%)	3,110,481	2.05	3.52	0.00	0.81	5.47
SDRET (%)	3,562,961	3.81	3.39	1.20	2.79	7.52
STR (%)	3,633,490	1.00	15.28	-14.90	0.00	16.82
MOM (%)	3,344,367	13.02	58.76	-45.21	5.56	71.43
ME ($ million)	3,634,603	1,327.86	7,025.86	5.49	70.86	1,878.65
BE/ME	2,383,096	0.00087	0.00079	0.0002	0.00067	0.0017
GP	2,463,191	0.32	0.29	0.04	0.29	0.69
ILLIQ	3,135,314	10.78	127.99	0.00	0.20	10.86

Table 2.1: Summary statistics. The table reports the number of stock-month observations (N), the mean (Mean), the standard deviation (Sd), and the 10th, 50th, and 90th percentiles (Pctile) of each variable. *SPREAD* is the monthly spread estimate. *SDRET* is the standard deviation of daily returns in the month. *ME* is the market capitalization in million dollars. *STR* is the stock's return in the previous month. *MOM* is the stock's return in the previous year, excluding the very last month. *BE/ME* is the book-to-market ratio as described in Fama & French (1992). *GP* is gross profitability as described in Novy-Marx (2013). *ILLIQ* is the Amihud (2002) illiquidity measure in the month. All variables are winsorized at the 1% level.

	Full Sample					Jan	Feb-Dec
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
SPREAD	21.62	18.26	22.15	24.48	10.57	91.23	15.30
	[6.89]	[5.60]	[6.58]	[7.35]	[2.77]	[6.30]	[4.95]
SDRET	−17.11	−17.65	−23.73	−20.52		54.89	−23.65
	[−6.14]	[−5.86]	[−7.83]	[−6.39]		[4.23]	[−8.72]
ME	−0.09	−0.10	−0.12			−0.61	−0.04
	[−3.56]	[−3.65]	[−4.43]			[−6.91]	[−1.63]
STR	−7.14	−6.99				−21.44	−5.84
	[−19.38]	[−18.04]				[−13.10]	[−16.84]
MOM	0.75					−2.25	1.03
	[3.86]					[−4.49]	[4.99]
N. Months	1140	1140	1140	1140	1140	95	1045
Min. Stocks	481	481	481	481	481	481	490
Med. Stocks	2192	2192	2192	2192	2192	2185	2192
Max. Stocks	6803	6803	6803	6803	6803	6754	6803

Table 2.2: Fama-MacBeth regressions of monthly excess returns against the spread and control variables. *SPREAD* is the average spread in the previous year. *SDRET* is the standard deviation of daily returns in the previous month. *ME* is the natural logarithm of market capitalization in the previous June. *STR* is the stock's return in the previous month. *MOM* is the stock's return in the previous year, excluding the very last month. All regressors are winsorized at the 1% level. The sample period is from January 1927 to December 2021. Columns 1 to 5 report results on the full sample period. Column 6 uses January months only. Column 7 uses months from February to December. The table reports the estimated coefficients multiplied by 100; t-statistics are reported in parentheses. The intercept is omitted from the table. The table further shows the total number of months and the minimum, median, and maximum number of stocks per month.

	1927–1962			1963–1999			2000–2021		
	All	Jan	Feb-Dec	All	Jan	Feb-Dec	All	Jan	Feb-Dec
SPREAD	15.55	87.95	8.97	23.48	89.82	17.45	28.43	98.98	22.02
	[3.22]	[3.98]	[1.89]	[5.63]	[4.44]	[4.32]	[3.35]	[2.52]	[2.61]
SDRET	−6.41	63.99	−12.81	−23.00	40.74	−28.80	−24.70	63.78	−32.74
	[−1.29]	[2.97]	[−2.59]	[−6.18]	[2.49]	[−7.89]	[−3.98]	[1.85]	[−5.68]
ME	−0.13	−0.55	−0.09	−0.07	−0.80	−0.01	−0.06	−0.40	−0.03
	[−2.48]	[−4.00]	[−1.63]	[−2.17]	[−5.60]	[−0.17]	[−1.41]	[−2.11]	[−0.67]
STR	−10.21	−25.36	−8.83	−7.40	−21.23	−6.14	−1.70	−15.37	−0.46
	[−14.68]	[−9.61]	[−12.99]	[−15.83]	[−8.51]	[−15.20]	[−2.53]	[−4.46]	[−0.76]
MOM	0.84	−2.62	1.15	1.15	−1.95	1.44	−0.06	−2.15	0.13
	[1.84]	[−3.15]	[2.36]	[7.19]	[−2.59]	[9.31]	[−0.22]	[−1.88]	[0.47]
N. Months	432	36	396	444	37	407	264	22	242
Min. Stocks	481	481	490	1093	1093	1098	3368	3376	3368
Med. Stocks	787	782	790	2418	2412	2419	3784	3860	3781
Max. Stocks	1091	1065	1091	6803	6754	6803	6025	6025	6004

Table 2.3: Fama-MacBeth regression of monthly excess returns against the spread and control variables in January and non-January months for three sub-periods. *SPREAD* is the average spread in the previous year. *SDRET* is the standard deviation of daily returns in the previous month. *ME* is the natural logarithm of market capitalization in the previous June. *STR* is the stock's return in the previous month. *MOM* is the stock's return in the previous year, excluding the very last month. All regressors are winsorized at the 1% level. The table reports the estimated coefficients multiplied by 100; t-statistics are reported in parentheses. The intercept is omitted from the table. The table further reports the total number of months and the minimum, median, and maximum number of stocks per month.

	Full Sample		1927–1962		1963–1999		2000–2021	
	EMC	IML	EMC	IML	EMC	IML	EMC	IML
				Panel A: Value Weighted				
				Summary Statistics				
Mean	0.34	0.30	0.47	0.49	0.36	0.26	0.09	0.05
St.dev	4.23	4.08	4.94	4.48	3.84	3.89	3.53	3.69
t-stat	2.72	2.46	1.97	2.26	1.99	1.40	0.43	0.22
				Alpha and Factor Loadings				
α	0.25	0.34	0.33	0.56	0.13	0.10	0.25	0.27
	[2.65]	[3.80]	[1.89]	[3.62]	[1.06]	[0.73]	[1.40]	[1.64]
β_{MKT}	−0.32	−0.45	−0.25	−0.44	−0.28	−0.33	−0.45	−0.54
	[−16.82]	[−24.86]	[−7.68]	[−15.16]	[−9.07]	[−10.17]	[−11.29]	[−14.25]
β_{SMB}	0.69	0.61	0.86	0.75	0.72	0.74	0.37	0.22
	[22.29]	[20.56]	[15.40]	[15.12]	[16.75]	[16.22]	[6.60]	[4.05]
β_{HML}	0.50	0.42	0.39	0.32	0.73	0.61	0.24	0.37
	[18.48]	[16.12]	[7.95]	[7.26]	[14.61]	[11.61]	[4.55]	[7.40]
				Panel B: Equally Weighted				
				Summary Statistics				
Mean	0.77	0.75	0.75	0.88	0.90	0.73	0.59	0.56
St.dev	4.28	4.43	4.93	5.03	3.83	4.08	3.86	3.92
t-stat	6.09	5.71	3.17	3.64	4.95	3.77	2.50	2.34
				Alpha and Factor Loadings				
α	0.67	0.69	0.55	0.76	0.73	0.57	0.84	0.82
	[6.76]	[6.85]	[3.67]	[5.10]	[5.24]	[3.85]	[4.11]	[4.07]
β_{MKT}	−0.29	−0.37	−0.25	−0.35	−0.29	−0.32	−0.47	−0.48
	[−14.69]	[−18.25]	[−9.09]	[−12.73]	[−8.26]	[−8.79]	[−10.20]	[−10.43]
β_{SMB}	0.63	0.66	0.84	0.97	0.60	0.69	0.21	0.03
	[19.59]	[19.89]	[17.38]	[19.98]	[12.34]	[13.46]	[3.23]	[0.41]
β_{HML}	0.53	0.54	0.57	0.50	0.63	0.63	−0.03	0.19
	[18.59]	[18.63]	[13.35]	[11.69]	[11.18]	[10.51]	[−0.49]	[3.09]

Table 2.4: Summary statistics, alpha and loadings on the Fama & French (1992) three-factor model for the expensive-minus-cheap (EMC) and illiq-minus-liquid (IML) factors. Mean and St.dev are the mean and standard deviation of the monthly returns (in %), respectively. t-stat is the t-statistic of the average return. α is the alpha with respect to the three-factor model (in %). The t-statistics on alpha and betas are given in parentheses. The first two columns report results in the full sample. Columns 3 to 8 report results for the three sub-samples. Panel A uses value-weighted factors. Panel B uses equally-weighted factors.

	Full Sample		1927–1962		1963-1999		2000-2021	
	VW	EW	VW	EW	VW	EW	VW	EW
				Panel A: IML				
α_{EMC}	0.06	0.02	0.26	0.19	−0.08	−0.16	−0.05	0.00
	[0.70]	[0.30]	[1.42]	[1.79]	[−1.02]	[−2.37]	[−0.39]	[0.03]
β_{EMC}	0.69	0.95	0.49	0.92	0.92	1.00	0.88	0.93
	[34.88]	[76.64]	[13.22]	[43.91]	[46.24]	[56.41]	[27.01]	[35.57]
Adj. R^2	0.52	0.84	0.29	0.82	0.83	0.88	0.73	0.83
				Panel B: EMC				
α_{IML}	0.12	0.11	0.18	−0.03	0.13	0.25	0.07	0.10
	[1.40]	[2.15]	[0.88]	[−0.28]	[1.75]	[3.94]	[0.59]	[1.02]
β_{IML}	0.75	0.88	0.59	0.89	0.90	0.88	0.83	0.89
	[34.88]	[76.64]	[13.22]	[43.91]	[46.24]	[56.41]	[27.01]	[35.57]
Adj. R^2	0.52	0.84	0.29	0.82	0.83	0.88	0.73	0.83

Table 2.5: Regressions of the illiq-minus-liquid (IML) factor on the expensive-minus-cheap (EMC) factor and vice-versa. Panel A regresses IML on EMC. Panel B regresses EMC on IML. α is given in %. The t-statistics are given in parentheses. The first two columns report results in the full sample for the value-weighted (VW) and equally-weighted (EW) factors. Columns 3 to 8 report results for the three sub-samples.

2.A Appendix

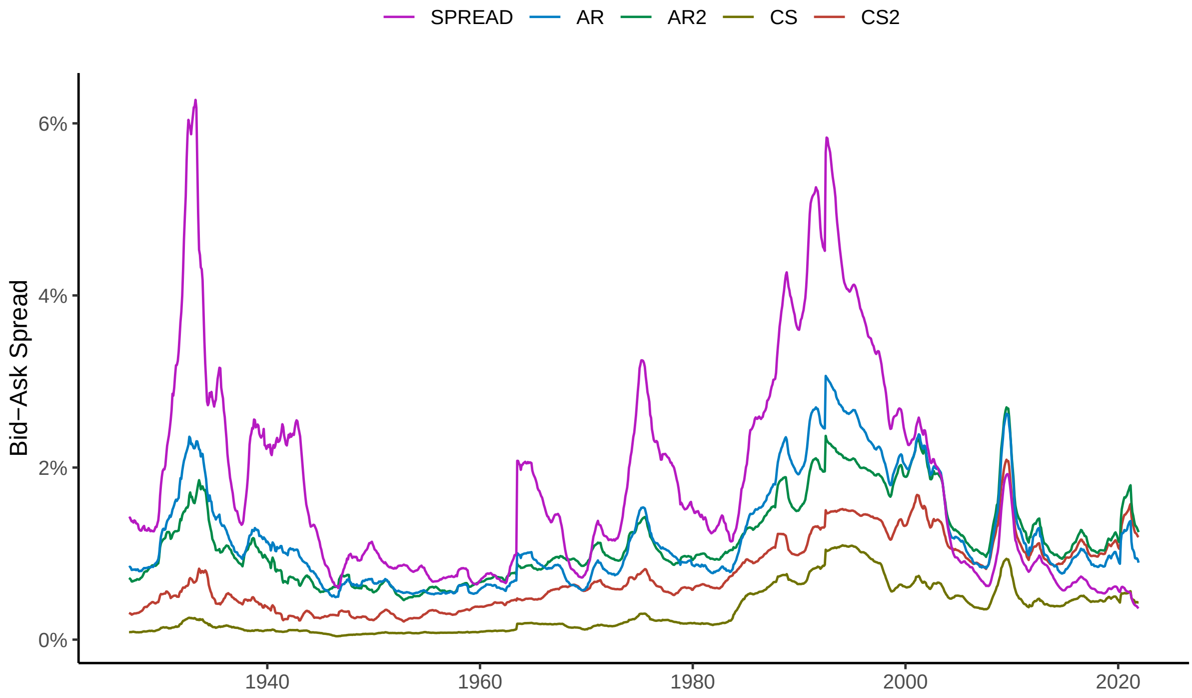


Figure 2.12: Time-series of bid-ask spread measures for an equally-weighted portfolio. *SPREAD* is the bid-ask spread used in this paper. *ROLL* is the bid-ask spread estimator in Roll (1984). *AR* and *AR2* are the monthly-adjusted and the daily-adjusted versions of the estimator in Abdi & Ranaldo (2017), respectively. *CS* and *CS2* are the monthly-adjusted and the daily-adjusted versions of the estimator in Corwin & Schultz (2012), adjusted for overnight returns. Each measure is computed monthly for each stock. Then, it is averaged using a 12-months rolling window for each stock. Finally, it is averaged (equally-weighted) across stocks.

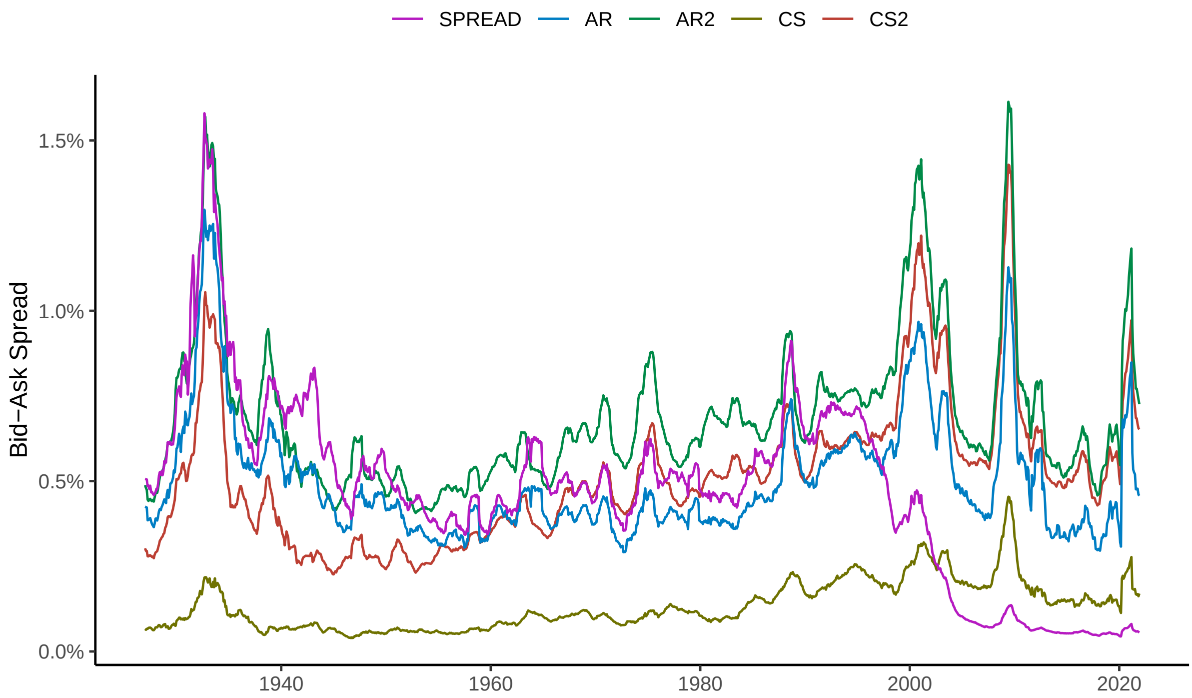


Figure 2.13: Time-series of bid-ask spread measures for a value-weighted portfolio. *SPREAD* is the bid-ask spread used in this paper. *ROLL* is the bid-ask spread estimator in Roll (1984). *AR* and *AR2* are the monthly-version and the daily-version of the estimator in Abdi & Ranaldo (2017), respectively. *CS* and *CS2* are the monthly-version and the daily-version of the estimator in Corwin & Schultz (2012), adjusted for overnight returns. Each measure is computed monthly for each stock. Then, it is averaged using a 12-months rolling window for each stock. Finally, it is averaged (value-weighted) across stocks.

	Full Sample							Jan	Feb-Dec
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
SPREAD	25.22	25.00	26.39	24.36	27.23	28.88	13.31	102.79	18.23
	[5.35]	[5.29]	[5.51]	[4.94]	[5.34]	[6.47]	[2.70]	[4.95]	[3.88]
SDRET	−18.71	−19.33	−21.23	−21.27	−25.15	−22.84		46.62	−24.59
	[−5.71]	[−5.88]	[−6.14]	[−5.60]	[−6.60]	[−5.61]		[2.76]	[−7.85]
ME	−0.02	−0.03	−0.06	−0.06	−0.08			−0.55	0.03
	[−0.74]	[−1.14]	[−2.35]	[−2.30]	[−2.67]			[−5.06]	[0.99]
STR	−5.39	−5.27	−5.02	−4.93				−17.42	−4.30
	[−14.04]	[−13.67]	[−12.85]	[−12.08]				[−8.81]	[−12.33]
MOM	0.56	0.59	0.62					−2.15	0.80
	[3.83]	[4.00]	[4.25]					[−3.15]	[5.61]
BE/ME	0.26	0.21						0.48	0.24
	[4.94]	[4.10]						[1.78]	[4.62]
GP	0.66							−0.55	0.77
	[6.05]							[−1.19]	[6.96]
N. Months	702	702	702	702	702	702	702	58	644
Min. Stocks	666	666	666	666	666	666	666	674	666
Med. Stocks	3094	3094	3094	3094	3094	3094	3094	3122	3089
Max. Stocks	5585	5585	5585	5585	5585	5585	5585	5373	5585

Table 2.6: Robustness test including book-to-market ratio and gross profitability. The table reports Fama-MacBeth regressions of monthly excess returns against the spread and control variables. *SPREAD* is the average spread in the previous year. *SDRET* is the standard deviation of daily returns in the previous month. *ME* is the natural logarithm of market capitalization in the previous June. *STR* is the stock's return in the previous month. *MOM* is the stock's return in the previous year, excluding the very last month. *BE/ME* is the natural logarithm of book-to-market as described in Fama & French (1992). *GP* is gross profitability as described in Novy-Marx (2013). All regressors are winsorized at the 1% level. The sample period is from July 1963 to December 2021. Columns 1 to 7 report results on the full sample period. Column 8 uses January months only. Column 9 uses months from February to December. The table reports the estimated coefficients multiplied by 100; t-statistics are reported in parentheses. The intercept is omitted from the table. The table further reports the total number of months and the minimum, median, and maximum number of stocks per month.

	Full Sample							Jan	Feb-Dec
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$SPREAD_{t-2}$	25.05	24.71	26.09	24.25	22.93	24.21	13.04	97.56	18.52
	[5.39]	[5.30]	[5.50]	[4.94]	[4.52]	[5.33]	[2.67]	[4.79]	[3.99]
$SDRET_{t-2}$	−17.10	−17.56	−19.29	−18.51	−16.78	−15.33		46.66	−22.84
	[−5.53]	[−5.64]	[−5.86]	[−5.03]	[−4.34]	[−3.70]		[3.17]	[−7.61]
ME	−0.02	−0.03	−0.06	−0.06	−0.06			−0.63	0.04
	[−0.68]	[−1.06]	[−2.27]	[−1.99]	[−2.15]			[−4.92]	[1.31]
STR	−5.70	−5.60	−5.38	−5.26				−16.14	−4.76
	[−15.43]	[−15.07]	[−14.30]	[−13.39]				[−8.39]	[−13.97]
MOM	0.59	0.62	0.65					−2.30	0.85
	[3.97]	[4.16]	[4.41]					[−3.24]	[5.90]
BE/ME	0.27	0.22						0.50	0.25
	[5.04]	[4.20]						[1.87]	[4.69]
GP	0.65							−0.63	0.77
	[6.01]							[−1.44]	[6.95]
N. Months	702	702	702	702	702	702	702	58	644
Min. Stocks	618	618	618	618	618	618	618	674	618
Med. Stocks	3090	3090	3090	3090	3090	3090	3090	3119	3088
Max. Stocks	5560	5560	5560	5560	5560	5560	5560	5370	5560

Table 2.7: Robustness test lagging regressors by one additional month. The table reports Fama-MacBeth regressions of monthly excess returns in month t against the spread and control variables. $SPREAD_{t-2}$ is the average spread in the previous 13 months, excluding the very last month. $SDRET_{t-2}$ is the standard deviation of daily returns in month $t-2$. *ME* is the natural logarithm of market capitalization in the previous June. *STR* is the stock's return in the previous month. *MOM* is the stock's return in the previous year, excluding the very last month. *BE/ME* is the natural logarithm of book-to-market as described in Fama & French (1992). *GP* is gross profitability as described in Novy-Marx (2013). All regressors are winsorized at the 1% level. The sample period is from July 1963 to December 2021. Columns 1 to 7 report results on the full sample period. Column 8 uses January months only. Column 9 uses months from February to December. The table reports the estimated coefficients multiplied by 100; t-statistics are reported in parentheses. The intercept is omitted from the table. The table further reports the total number of months and the minimum, median, and maximum number of stocks per month.

Chapter 3

A Theory of Price Formation in Financial Markets

Mainstream asset pricing is based on the idea that the asset price is the discounted value of future payoffs. The development of economic theories that explain the origin of financial market fluctuations has proceeded along two complementary paths that focus on either the payoffs or the discount rate. The first school of thought views the market as an efficient machinery that quickly and objectively estimates futures payoffs and discounts them at a certain rate to output the current price (Fama, 1970). In this view, prices change because of the release of new information that objectively changes the estimation of future payoffs. Although theoretically appealing, this approach has faced inconsistencies with theoretical arguments and empirical data. In particular, markets cannot be efficient if acquiring information is costly (Grossman & Stiglitz, 1980), and asset prices are too volatile relative to what can be attributed to new information about future payoffs (Shiller, 1981). The second school of thought has provided an alternative in recent years, in which markets are driven by investor behaviors and cognitive biases (Shiller, 2003). In this view, prices change because of preference shifts that subjectively change the discount rate. This approach has contributed to improving the understanding of asset prices and revealed several important empirical regularities (Shiller, 2014). Despite these advances, little is known about the actual mechanism through which the price form. Indeed, both theories predict that prices form without trading and volume is essentially zero or due to irrational behaviours (Milgrom & Stokey, 1982). From this perspective, the astronomical amount of trading volume observed in financial markets would be only explainable by human folly. This is an important puzzle.[1]

In this paper, I study price formation starting from the fact that trading does happen, and investigate its mechanical impact on prices. The theory proposed here consists of three fundamental propositions about price impact and price formation. Proposition I introduces a market clearing condition that encodes how orders are matched in modern electronic markets. This condition states that the quantity exchanged in a trade equals the integrated density of the book between zero and the peak impact. Proposition II introduces a condition of fair pricing such that both parties involved in a trade earn zero profits in the transaction. This condition implies that the peak impact relaxes to a non-zero permanent impact. Finally, Proposition III states that asset prices are the result of the accumulation of price impacts. The theory proposed here presents a mechanical view of financial markets where trading *itself* is the source of price fluctuations.

To evaluate the validity of the theory, I develop several predictions about the joint relationships among trading volume, price impact, trading frequency, bid-ask spreads, order imbalance, market depth, and volatility that follow from Propositions I, II, and III

[1] On this matter, see "The Grumpy Economist: Volume and Information" by J. Cochrane: `https://johnhcochrane.blogspot.com/2016/10/volume-and-information.html`

under general conditions. First, price impact increases approximately with the square root of trading volume, with permanent impact relaxing to two-thirds of peak impact. Additionally, the impact is smaller when market depth is larger. Second, financial markets exhibit two regimes. In the first regime, an investor who buys (sells) moves prices up (down) and decreases (increases) future expected returns. In the second regime, expected returns diverge. Third, market makers post more and larger sell orders when market takers post more and larger buy orders, and vice versa. Fourth, volatility rises with larger trading volumes and decreases with larger market depths. It also depends on the order imbalance with strong asymmetric effects. Generally, volatility signature plots decrease but may increase when price impacts are positively correlated. Fifth, a trade of fixed size has a larger impact if executed during periods of high volatility and low volume. Sixth, bid-ask spreads are proportional to volatility per unit of trade. Seventh, assets that are traded less frequently exhibit greater autocorrelation in returns. Finally, trading invariants are not invariant and a more general invariant is proposed.

To gather preliminary evidence, I survey previous findings in the literature and provide novel empirical results. The analyses are performed using the CRSP U.S. Stock Database, which provides low-frequency data dating back to 1926. Analyses that require high-frequency data are left for future work.

This paper mainly relates to the literature that aims at understanding the origins of financial market fluctuations (Fama, 1970; Shiller, 1981, 2003, 2014; Gabaix & Koijen, 2021). It also contributes to several strands of literature that are discussed in the main body of the paper.

This work is structured as follows. Section 3.1 presents the model. Section 3.2 develops the theoretical predictions. Section 3.3 presents the empirical results. Finally, Section 3.4 concludes. All proofs and calculations are reported in the appendix.

3.1 Model

Modern financial markets implement trading via an electronic order-matching system. Traders submit their orders to the system, which collects this information in an *order book*. Buyers post bids that state the maximum price at which they are willing to buy, while sellers post offers that state the minimum price at which they are willing to sell. The highest price of buy orders is called *bid* price, while the lowest price of sell orders is called *ask* price. The average of the two is the *midpoint*. When a new order is received, the system updates the order book. A new buy (sell) order is matched if its price is greater (less) than or equal to the ask (bid) price. If the order is matched, a certain quantity is traded at the ask price (for buys) or bid price (for sells) and removed from the book. Any remaining quantity is executed at the next most favorable price until the order is filled or a limit price is reached. In general, only a portion of the total order size is executed. I denote the executed quantity with Q, and the convention is that Q is positive for buys and negative for sells.

In this work, I model an order book with continuous prices. At each instant of time, a function $\rho(x) \geq 0$ gives the order book density at the price that is at distance x from the midpoint. By construction, $\rho(x)$ is the density of sell orders for $x > 0$ (ask side); it is the density of buy orders for $x < 0$ (bid side); and $\rho(0) = 0$ as there are no buy or sell orders in the midpoint. An illustration is given in Figure 3.1.

[Insert Figure 3.1 about here.]

Any trade has a price impact. Indeed, a first portion of the quantity Q is traded near the midpoint and removed from the book. This process iterates by trading at prices incrementally further away from the midpoint until Q is executed and a final price is reached. The difference between the final price and the initial midpoint is the peak impact Δ.

Proposition I (Market clearing). *For any trade, the exchanged quantity Q is equal to the integrated density of the book between zero and the peak impact Δ.*

$$\boxed{Q = \int_0^\Delta \rho(x)dx} \tag{3.1}$$

The peak impact generally consists of a permanent component δ and a temporary component $\Delta - \delta$ which vanishes in the long run. To quantify the components of the peak impact, I consider the profits earned by both parties involved in the transaction. The trade initiator, who submitted the order that triggered the transaction, is called *market taker*, and its counterpart is called *market maker*. Let $dQ = \rho(x)dx$ be the infinitesimal quantity traded at a price x from the initial midpoint. In equilibrium, the infinitesimal profit earned by the market taker in the transaction is $d\pi = (\delta - x)dQ$, and the total profit is $\pi = \int_0^\Delta d\pi$. By symmetry, the market maker earns $-\pi$. For the price to be fair, the maker's profit must equal the taker's profit. Otherwise, competition among market participants would reduce the profits of makers or takers, whichever is more profitable, and increase the profits of the counterpart. Thus, a unique equilibrium exists where both parties earn $\pi = 0$.

Proposition II (Fair pricing). *The peak impact Δ relaxes to the permanent impact δ, such that makers and takers earn zero profits in the transaction.*

$$\boxed{\int_0^\Delta (x - \delta)\rho(x)dx = 0} \tag{3.2}$$

The dynamics of price impacts is represented by an impact function $\Delta_i(n)$, defined for $n \geq 0$, which gives the price impact of the i-th trade after n subsequent trades. The impact of the i-th trade after it is executed is its peak impact $\Delta_i(0) = \Delta_i$. In the long run, its impact relaxes to the permanent impact $\lim_{n\to\infty} \Delta_i(n) = \delta_i$. In the model, the asset price is determined by the accumulation of price impacts.

Proposition III (Price formation). *The asset price is determined by an initial reference price P_0 and the impact functions of all previous trades.*

$$\boxed{P_j = P_0 + \sum_{i=1}^{j} \Delta_i(j - i)} \tag{3.3}$$

Equations (3.1)–(3.3) may be solved with any particular choice of the order book density and impact functions. This paper develops several predictions that these equations imply under general conditions.

3.1.1 Trading Around the Midpoint

For trades executed near the midpoint ($x = 0$), the order book density admits the following linearization:

$$\rho(x) = \begin{cases} \alpha x & \text{if } x > 0 \\ \beta x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases} \tag{3.4}$$

where α and β are the right-hand and left-hand derivatives of the density, obtained by approaching the midpoint from the ask and bid side of the book, respectively. Given that the density is positive, it follows that $\alpha > 0$ and $\beta < 0$ (see Figure 3.1). Since books with steeper slopes are denser, α and β are measures of market depth.

3.1.2 Impact Functions

The impact functions have zero mean to avoid upward or downward trends in the price process:

$$\mathbb{E}[\Delta_i(n)] = 0\,. \tag{3.5}$$

3.2 Theoretical Predictions

This section develops several predictions obtained by solving Equations (3.1)–(3.3) with the linearization in (3.4) and the impact functions in (3.5).

3.2.1 Price Impact

In the seminal work by Kyle (1985), price impact is a linear function of trade size. The prediction of linearity is reinforced by Huberman & Stanzl (2004) who show that, under certain conditions, price impact must be linear to prevent quasi-arbitrage. However, empirical data contradict this prediction and empirical studies consistently find concave impact. Farmer et al. (2013) develop a theory predicting that price impact should increase as the square root of traded volume, with permanent impact relaxing to two-thirds of peak impact. These predictions are obtained under a condition of fair pricing and by assuming a particular distribution for order sizes. The main objection to this model is the importance given to the order size distribution, which appears to be less general than the model's predictions (Zarinelli et al., 2015). Indeed, a large number of empirical studies have concluded that price impact increases approximately with the square root of trading volume, independently of the asset class, time period, style of trading, and market structure. Lillo et al. (2003) provides evidence for small transactions in high-capitalization stocks. Several studies on metaorders' impact provide evidence in a variety of stock markets (Almgren et al., 2005; Moro et al., 2009; Gomes & Waelbroeck, 2015; Bacry et al., 2015; Brokmann et al., 2015; Said et al., 2017; Bucci et al., 2018), in option markets (Tóth et al., 2016; Said et al., 2021), for futures contracts (Tóth et al., 2011), and bitcoin (Donier & Bonart, 2015). Moreover, Moro et al. (2009); Zarinelli et al. (2015) and Said et al. (2017) find that permanent impact relaxes to two-thirds of peak impact after one day, while (Bucci et al., 2018) find that the decay continues the next days and converges to a value of about one-third after 50 days.

Here, price impact is studied by solving Equation (3.1) with the density in (3.4). The peak impact for the i-th trade is:

$$\Delta_i^{(buy)} = +\sqrt{\frac{2Q_i}{\alpha_i}} \quad \text{or} \quad \Delta_i^{(sell)} = -\sqrt{\frac{2Q_i}{\beta_i}}\,, \tag{3.6}$$

where $\Delta_i^{(buy)}$ is the peak impact if the trade is buyer-initiated (executed on the ask side), and $\Delta^{(sell)}$ is the peak impact if the trade is seller-initiated (executed on the bid side). This result shows that buyer-initiated trades move prices upward and seller-initiated trades move prices downward. The peak impact increases as the square root of traded volume and it is inversely proportional to the market depth, which is represented by α_i for buyer-initiated trades and β_i for seller-initiated trades. Furthermore, by solving Equation (3.2) with the density in (3.4), the permanent component of price impact is two-thirds of peak impact:

$$\delta_i = \frac{2}{3}\Delta_i\,. \tag{3.7}$$

These results can be summarized as follows:

$$\Delta_i(n) = \Delta_i f(n) \quad \text{with} \quad f(n) : 1 \downarrow 2/3\,, \tag{3.8}$$

where Δ_i is the peak impact given in Equation (3.6) and $f(n)$ is a function decreasing from $f(0) = 1$ to $\lim_{n\to\infty} f(n) = 2/3$. Taken together, these results say that peak impact increases as the square root of traded volume, with permanent impact relaxing to two-thirds of peak impact. An illustration is given in Figure 3.2.

[Insert Figure 3.2 about here.]

Furthermore, the model implies that these predictions should hold more accurately for smaller trades executed near the midpoint, as they better satisfy the first-order approximation of the density in Section 3.1.1. For instance, the predictions should hold accurately for small transactions in liquid assets and for metaorders, which are split in small pieces by construction. On the other hand, one should expect significant deviations for larger trades executed deeper in the book, such as large transactions in illiquid markets. In support of this idea, deviations are observed for larger transactions in smaller stocks (Lillo et al., 2003), larger-tick futures contracts (Tóth et al., 2011) and larger metaorders (Bershova & Rakhlin, 2013; Zarinelli et al., 2015).

One limitation of Equations (3.6)–(3.8) is that they formally hold for single transactions, while most of the empirical evidence is on aggregate transactions (metaorders). The generalization of the model's predictions to metaorders' impact is left for future work.

3.2.2 Return Predictability

Consider an investor who aims to generate profits by predicting returns. The investor buys a quantity Q_j when the expected return is positive and sells otherwise. However, the expected return is endogenously determined by the impact that the investor generates to trade. Indeed, according to Equation (3.3), the return after n trades is given by:

$$P_{j+n} - P_j = \underbrace{\sum_{i=1}^{j} \Delta_i(j+n-i) - \Delta_i(j-i)}_{\mu_j(n)} + \underbrace{\sum_{i=j+1}^{j+n} \Delta_i(j+n-i)}_{w_j(n)}, \tag{3.9}$$

which includes the component $\Delta_j(n) - \Delta_j(0) = -\Delta_j(1-f(n))$ generated by the investor to trade the quantity Q_j. As $f(n)$ is smaller than 1, the component has always the opposite sign of the peak impact Δ_j. If the investor buys ($\Delta_j > 0$), trading decreases the expected return. If the investor sells ($\Delta_j < 0$), trading increases the expected return. Thus, the investor faces a tradeoff between trading a large quantity and having a small impact on expected returns. To choose the optimal quantity to trade, the investor maximizes the expected profits:[2]

$$\max_{Q_j} \; Q_j \mathbb{E}_j[P_{j+n} - P_j], \tag{3.10}$$

where $\mathbb{E}_j$ denotes the expectation conditional on past information. The expected return can be decomposed in two parts:

$$\mathbb{E}_j[P_{j+n} - P_j] = \mathbb{E}_j[P_{j+n} - P_j \mid Q_j = 0] - \Delta_j\big(1 - f(n) - g_j'(n)\big). \tag{3.11}$$

The first part is independent from the traded quantity and corresponds to the return expected if no trade takes place ($Q_j = 0$). The second part takes into account the dependence on Q_j through its peak impact Δ_j. The term $g_j'(n)$ is the derivative of the expected value of $w_j(n)$ in Equation (3.9) with respect to Δ_j. This term measures the dependence of future impacts on the traded quantity Q_j. If $g_j'(n) < 0$, then buying (selling) reduces future buying (selling) pressure. If $g_j'(n) > 0$, then buying (selling) increases future buying (selling) pressure.

The optimal trading quantity is computed by replacing Δ_j in Equation (3.11) with the peak impacts in Equation (3.6) and solving Equation (3.10) for Q_j. The solution is that Q_j is finite only if $g_j'(n) < 1 - f(n)$. Otherwise, Equation (3.11) shows that the expected return increases when the investor buys and decreases when the investor sells. Thus, the investor does not face any tradeoff and the optimal strategy is to buy or sell the maximum possible quantity. Using the optimal quantity into Equation (3.11) gives the following expected return.

$$\mathbb{E}_j[P_{j+n} - P_j] = \begin{cases} \frac{1}{3}\mathbb{E}_j[P_{j+n} - P_j \mid Q_j = 0] & \text{for} \quad g_j'(n) < 1 - f(n) \\ \pm\infty & \text{for} \quad g_j'(n) \geq 1 - f(n) \end{cases} \tag{3.12}$$

Equation (3.12) shows that the expected return exhibits two regimes. In the first regime, the expected return after trading is a fraction of the return expected before trading. An investor who buys (sells) moves prices up (down) and decreases (increases) future expected returns. After a certain amount of trades, returns are incorporated into

[2]The investor may also maximize the expected profits adjusted for some risk measure. However, as long as risk does not depend on the traded quantity Q_j, this is equivalent to the problem in Equation (3.10).

prices and expected returns shrink to zero. Prices are thus difficult to predict because predictability vanishes fast, as investors acting first are those rewarded most. From this perspective, market efficiency, here intended as the apparent lack of return predictability, should not be seen as a property of the market, but rather as a measure of skill of the market participants. Indeed, Equation (3.9) shows that returns are intrinsically predictable, not only by predicting future impacts, but also from the knowledge of past impacts.[3] With little trading, returns would be predictable and markets would be regarded as inefficient. But skilled investors exploit such predictability by trading to generate profits. As a result, trading volume increases, predictability declines, and markets appear more efficient. This view is in line with the empirical finding of Chordia et al. (2008) that more liquid markets are more efficient, and of McLean & Pontiff (2016) that efficiency improves as investors become more skilled.

The second regime is radically different. Here, an investor who buys (sells) moves prices up (down) and increases (decreases) future expected returns. This is the case when a buy (sell) order is interpreted by future trades as a buy (sell) signal and generates future price impacts in the same direction. Predictability spikes and expected returns diverge. This regime is less frequent than the first one because $g'_j(n) = 0 < 1 - f(n)$ in the typical case where price impacts are uncorrelated. Moreover, it becomes less frequent for longer horizons because the threshold to exceed $1 - f(n)$ increases for larger n. Thus, this regime resembles bubbles and crashes, and the model makes the reasonable prediction that flash crashes are more frequent than economic recessions. Importantly, this regime does not arise from behavioural biases or irrationality. Instead, it represent a very rational response to self-fulfilling prophecies (Blanchard & Watson, 1982). By predicting that buy or sell trades will amplify price movements in the same direction, the optimal strategy is to enter or quit the market with large quantities as quickly as possible. This result may help to explain why predictability spikes (Cujean & Hasler, 2017) and liquidity evaporates (Nagel, 2012) during periods of financial turmoil, and to better rationalize market inefficiencies.

3.2.3 Order Imbalance

Chordia et al. (2005) reconcile the persistence in order imbalances with the lack of predictability in returns. The problem is that, on one side, there is a high level of predictability in order imbalances and investors tend to continue buying or selling for an extended period. On the other side, returns are uncorrelated and continuous buy or sell pressure does not seem to predict price movements. This suggests that "some astute investors must be correctly forecasting continued price pressure from order imbalances and conducting countervailing trades [...] sufficient to remove all serial dependence in returns, which would otherwise be induced by the continuing procession of order imbalances." (Chordia et al., 2005, p. 273). They show that sophisticated investors react to order imbalances by undertaking enough countervailing trades to remove serial dependence in returns.

[3] The component $\mu_j(n)$ depends on past information as the impact functions are determined for all $i < j$ at the time when the j-th trade is executed and they can be evaluated at future trades. On the other hand, the component $w_j(n)$ depends on future information because the impact functions for $i > j$ are not determined at the time when the j-th trade is executed.

Here, the order-flow imbalance $\phi_t \in [-1, 1]$ is defined as the difference between the probability of buyer-initiated trades minus the probability of seller-initiated trades in period t. In the same period, the average market depth γ_t and the order-book imbalance θ_t are defined by:

$$\gamma_t = \frac{\alpha_t - \beta_t}{2} \quad \text{and} \quad \theta_t = \frac{\beta_t + \alpha_t}{\beta_t - \alpha_t}\,, \tag{3.13}$$

where α_t is the average α_i of buys and β_t is the average β_i of sells. The order-book imbalance $\theta_t \in [-1, 1]$ is positive when the book is buy heavy ($|\beta| > |\alpha|$) and it is negative when the book is sell heavy ($|\alpha| > |\beta|$). The relation between order-book imbalance and order-flow imbalance is obtained by solving Equation (3.5) with the impacts in Equation (3.8):

$$\frac{1-\theta_t}{1+\theta_t} = \left(\frac{1+\phi_t}{1-\phi_t}\right)^2 \left|\frac{Q_t^{(buy)}}{Q_t^{(sell)}}\right|\,, \tag{3.14}$$

where $Q_t^{(buy)}$ and $Q_t^{(sell)}$ are the average trade sizes of buys and sells, respectively. This relation states that the book is buy heavy ($\theta_t > 0$) when the probability of sells is higher ($\phi_t < 0$), and it is sell heavy ($\theta_t < 0$) when the probability of buys is higher ($\phi_t > 0$). Similarly, the book is buy heavy when the average size of sells increases ($|Q_t^{(sell)}| > |Q_t^{(buy)}|$), and it is sell heavy when the average size of buys increases ($|Q_t^{(buy)}| > |Q_t^{(sell)}|$). Thus, order-book imbalance (market makers) counteracts order-flow (market takers) imbalance, consistent with the empirical evidence that investors react to order imbalances by undertaking countervailing trades to remove serial dependence in returns.

3.2.4 Volatility

A large body of literature has documented the empirical relation between volatility and trading volume. Karpoff (1987) surveys 18 distinct works that document this relation in a variety of markets. Most of these studies document a positive relation between volume and the square of price changes. However, linking volatility to volume does not extract all information. For instance, Bessembinder & Seguin (1993) find that volatility is negatively related to open interest, which is used as a proxy for market depth. Jones et al. (1994) find that the volatility-volume relation is driven by the trading frequency, and Chan & Fong (2000) document the role of the order imbalance.

In this work, there is no exogenous volatility. Volatility is endogenously determined by price impacts. The expression for the volatility is obtained by computing the variance of price impacts in Equation (3.3):

$$\sigma_t^2 = \mathbb{V}ar[P_{j+N_t} - P_j] = \frac{2V_t\lambda_t^2 f_t^2}{\gamma_t(1+\theta_t\phi_t)}\,, \tag{3.15}$$

where V_t is the total volume traded in the period, N_t is the number of trades in the period, γ_t is the average market depth, θ_t and ϕ_t are the order-book and order-flow imbalances, respectively, and:

$$\lambda_t^2 = \frac{\mathbb{V}ar\left[\sum_{i=j+1}^{j+N_t} \Delta_i(j+N_t-i)\right]}{\sum_{i=j+1}^{j+N_t} \mathbb{V}ar\left[\Delta_i(j+N_t-i)\right]}\,, \quad f_t^2 = \frac{1}{N_t}\sum_{i=j+1}^{j+N_t} f^2(j+N_t-i)\,. \tag{3.16}$$

Equation (3.15) shows that volatility is linked to volume. Trading larger volumes creates larger price impacts which increase volatility. In turn, larger market depths reduce the impact of trades and decrease volatility. The order imbalances have a strong asymmetric effect. Volatility reduces by a small factor when the imbalances have the same sign, but it increases by a factor that can be arbitrarily large when the imbalances have opposite signs. This is the scenario where the probability of buys (sells) is high and the book is sell (buy) heavy. In this case, the more frequent trades have a smaller impact as they face a larger market depth, but the less frequent trades in the opposite direction have a larger impact as they face a smaller market depth and their trade size is typically larger.[4] Strong reversal are thus expected in this scenario and volatility spikes. Volatility also depends on the number of trades through the quantities defined in Equation (3.16). Here, f_t is the root-mean-square f within the period and decreases as N_t increases. This implies that volatility per unit of time is not the same when measured at different sampling frequencies. High-frequency estimates are higher than low-frequency estimates because the time periods contain fewer trades and price impacts do not have enough time to relax. In other words, low-frequency estimates of volatility consider the permanent component of price impacts while high-frequency estimates also include the transitory component. The term λ_t^2 is the ratio between the variance of the sum and the sum of the variances of all price impacts in the period. This term is equal to 1 when price impacts are uncorrelated, it decreases to 0 when they are negatively correlated, and it increases to N_t when they are positively correlated. In the latter case, volatility increases with lower sampling frequencies and high-frequency estimates may be lower than low-frequency estimates. In general, the product $\lambda_t^2 \times f_t^2$ may either increase or decrease with the number of trades, depending on the correlation structure of price impacts. The result is that volatility signature plots (Andersen et al., 2000) may be either increasing or decreasing, but certainly not flat as a random walk model would imply.

3.2.5 The Square-Root Law

Price impacts can be linked to volatility using Equations (3.8) and (3.15):

$$\Delta_i(n) = Y_i(n) \times \sigma_t \sqrt{\frac{|Q_i|}{V_t}}\,. \tag{3.17}$$

This expression resembles the square-root law, which is well established empirically (see Bouchaud et al., 2018, and references therein) and also discussed in the handbook by Grinold & Kahn (2000) as a practical way for asset managers to estimate market impact. This law states that the impact of a trade (metaorder) is proportional to volatility and to the square root of the ratio between trade size and total volume. In other words, a trade of fixed size has a larger impact if executed in periods when volatility is high and volume is low. One important difference is that here Q_i is the size of a single transaction. The link with aggregate transactions requires to understand how Equation (3.8) generalizes to metaorders' impact and is left for future work.

[4] Set $\theta_t = -\phi_t$ in Equation (3.14). This implies $|Q_t^{(sell)}| > |Q_t^{(buy)}|$ for $\phi_t > 0$, and vice-versa for $\phi_t < 0$.

The derivation of Equation (3.17) also provides an explicit formula for the coefficient of proportionality:

$$Y_i(n) = \begin{cases} \dfrac{f(n)}{\lambda_t f_t}\sqrt{\dfrac{\gamma_t(1+\theta_t\phi_t)}{\alpha_i}} & \text{for buys} \\ \dfrac{-f(n)}{\lambda_t f_t}\sqrt{\dfrac{\gamma_t(1+\theta_t\phi_t)}{-\beta_i}} & \text{for sells} \end{cases} \tag{3.18}$$

The coefficient incorporates the dependence on $f(n)$: It is higher for peak impacts ($n = 0$) and reduces by a factor 2/3 for permanent impacts ($n \to \infty$). Considering a typical scenario where $f(n)$ is approximately equal to its root-mean-square value, price impacts are uncorrelated, there is no order imbalance, and market depths are constant, one has that $f(n) = f_t$, $\lambda_t = 1$, $\theta_t = 0$, and $\alpha_t = \alpha_i = -\beta_t = -\beta_i = \gamma_t$. Thus, $|Y_i(n)| = 1$. In other words, the coefficient of proportionality is expected to be of order unity and $|Y_i(n)| \approx 1$ under general conditions, in agreement with several empirical studies (Tóth et al., 2011; Donier & Bonart, 2015; Tóth et al., 2016; Bouchaud et al., 2018).

3.2.6 Bid-Ask Spreads

For each trade, the effective bid-ask spread is defined as $S_i = 2D_i(P_i - P_i^{(mid)})$, where P_i is the price after the trade is executed, $P_i^{(mid)}$ is the midpoint price before the trade is executed, and D_i is an indicator variable that equals +1 if the trade is a buy and −1 if the trade is a sell (Holden & Jacobsen, 2014). The Lee & Ready (1991) algorithm classifies a trade that occurs above (below) the midpoint as a buy (sell). Thus, the effective spread is $S_i = 2|P_i - P_i^{(mid)}| = 2|\Delta_i|$, where $|\Delta_i|$ is the modulus of the peak impact. Computing the variance of price impacts in function of the spread establishes the following relation between the spread and volatility:

$$S_t = \frac{2}{\lambda_t f_t}\frac{\sigma_t}{\sqrt{N_t}} \tag{3.19}$$

This equation states that the spread is proportional to volatility per unit of trade and provides an expression for the coefficient of proportionality. Under the assumption that price impacts are weakly correlated ($\lambda_t \approx 1$), the coefficient of proportionality is between 2 and 3 as f_t ranges between 1 and 2/3.

The linear relation between bid-ask spreads and volatility per trade was first noted by Zumbach (2004) and further developed by Wyart et al. (2008). These studies report a coefficient of proportionality between 1 and 4, in good agreement with the model's predictions. An empirical extension to these studies is detailed in Section 3.3.1.

3.2.7 Trading Frequency

This is the first paper to document the possibility to estimate the trading frequency from price data. Indeed, Equation (3.19) can be inverted to compute the number of trades from the bid-ask spread and volatility:

$$N_t = \left(\frac{2}{\lambda_t f_t}\frac{\sigma_t}{S_t}\right)^2 . \tag{3.20}$$

For instance, the number of trades per day can be computed from the ratio between the volatility of daily returns and the bid-ask spread. Moreover, several studies have shown that the bid-ask spread can be estimated from transaction prices (Roll, 1984; Corwin & Schultz, 2012; Abdi & Ranaldo, 2017; Ardia et al., 2022). Thus, Equation (3.20) implies that prices contain all the information needed to determine the trading frequency. In particular, Roll (1984) shows that the spread is linked to the autocovariance of returns, such that $S_t = 2\sqrt{-\mathbb{C}\mathrm{ov}[R_t, R_{t-1}]}$ where R_t is the return in period t and R_{t-1} is the return in the previous period. Using this expression for the spread in Equation (3.20), one obtains that the number of trades is proportional to the inverse of the autocorrelation of returns:

$$\mathbb{C}\mathrm{or}[R_t, R_{t-1}] = \frac{-1}{\lambda_t^2 f_t^2 N_t}\,. \tag{3.21}$$

This relation is of conceptual interest in that it shows that assets traded less frequently exhibit a larger autocorrelation in returns. Furthermore, the autocorrelation is negative and it vanishes in the limit where the asset is traded a large amount of times. In practice, however, the Roll (1984) method is affected by a large estimation variance and it is preferrable to estimate the spread with more efficient methods (e.g, Ardia et al., 2022) to compute the number of trades via Equation (3.20), as demonstrated in Section 3.3.2.

3.2.8 Market Microstructure Invariance

Kyle & Obizhaeva (2016) propose an empirical hypothesis that market microstructure characteristics become constants when measured in units of trades instead of units of time. This hypothesis predicts that the quantity $\sigma_t V_t / N_t^{3/2}$ is invariant across assets and across time.[5] However, a direct calculation using Equations (3.15) and (3.19) shows that:

$$\frac{\sigma_t V_t}{N_t^{3/2}} = \frac{\lambda_t f_t \gamma_t (1 + \theta_t \phi_t) S_t^3}{16}\,. \tag{3.22}$$

Equation (3.22) reveals that the invariance hypothesis holds only in special circumstances because the quantity $\sigma_t V_t / N_t^{3/2}$ that should be invariant across assets and across time turns out to depend on a number of factors that vary in the cross-section and in the time-series. Specifically, it depends on the spread (S_t), market depth (γ_t), order imbalance (θ_t and ϕ_t), correlation structure of price impacts (λ_t), and sampling frequency (f_t). A more general invariant is thus:

$$I := \frac{16 \sigma_t V_t}{\lambda_t f_t \gamma_t (1 + \theta_t \phi_t) S_t^3 N_t^{3/2}} = 1\,. \tag{3.23}$$

Notice that I is a dimensionless variable. Indeed, σ_t and S_t are measured in units of price, V_t is measured in units of shares, γ_t is measured in units of shares over price squared, and λ_t, f_t, θ_t, ϕ_t, and N_t are dimensionless numbers.

[5] Here, $\sigma_t = \sigma_\$$ is the dollar volatility as defined in Equation (3.15). Kyle & Obizhaeva (2016) use the percent volatility $\sigma_\%$ and the prediction is that $\sigma_\% \times P_t \times V_t \,/\, N_t^{3/2}$ is invariant. The two notations are equivalent as $\sigma_\$ = \sigma_\% \times P_t$.

This more general invariant also scales with $N_t^{3/2}$ but includes several other factors. The 3/2 scaling was empirically confirmed by Kyle & Obizhaeva (2016) using a proprietary data set of portfolio transitions for U.S. clients, and by several other studies in a variety of different markets (Bae et al., 2016; Benzaquen et al., 2016; Andersen et al., 2018; Pohl et al., 2020; Kyle et al., 2020). However, while the 3/2 scaling is very robust, the empirical data revealed that the invariant $\sigma_t V_t / N_t^{3/2}$ is actually quite far from being invariant as it varies across assets and across time. The more general invariant proposed in Equation (3.23), which includes the bid-ask spread and a number of other factors, needs to be tested with high-frequency data and is left for future work. Still, encouraging evidence in this direction can be found in Benzaquen et al. (2016) and Bucci et al. (2020), who find that normalizing the invariance measure by Kyle & Obizhaeva (2016) for the bid-ask spread significantly improves the fit with the empirical data. Another important difference is that N_t in Equation (3.23) represents the number of trades, while in the Kyle & Obizhaeva (2016) invariant it is interpreted as the number of "bets" (metaorders).

3.3 Empirical Results

This section provides novel empirical evidence for the model's predictions and it illustrates potential applications. The analyses are performed using the CRSP U.S. Stock Database, which provides low-frequency data dating back to 1926. Analyses that require high-frequency data are left for future work.

3.3.1 Bid-Ask Spread and Volatility per Trade

The linear relation between bid-ask spreads and volatility per trade was first noted by Zumbach (2004) and further developed by Wyart et al. (2008). The analyses are performed using high-frequency data for 92 companies of the FTSE 100 in the sample period 2001–2002 (Zumbach, 2004), for 68 stocks of the Paris Stock Exchange in 2002, for 155 stocks of the NYSE in 2005, and for Futures contracts in 2005 (Wyart et al., 2008).

Here I extend the analysis using the CRSP U.S. Stock Database. CRSP provides the number of trades made on each date for a security listed in the NASDAQ Stock Market since 1982.[6] To compute the volatility per trade, I compute the standard deviation of daily returns in each month and divide by the square root of the average number of daily trades within the month. For each month, effective bid-ask spreads are computed from daily prices using the efficient estimator proposed by Ardia et al. (2022). This estimator is an improved and considerably more accurate version of the popular Roll (1984) estimator and other estimators from daily prices (Corwin & Schultz, 2012; Abdi & Ranaldo, 2017). These estimates of effective spreads are different from the the end-of-day quoted spread provided by CRSP. Indeed, quoted spreads can overstate effective spreads by up to 100% (Huang & Stoll, 1994; Petersen & Fialkowski, 1994; Bessembinder & Kaufman, 1997; Bacidorea et al., 2003), due to dealers offering a better price than the quotes, also known

[6] The daily number of trades is reported for all securities listed on The NASDAQ National Market since November 1, 1982, and all NASDAQ securities since June 15, 1992. Due to lack of sources, the data are missing for 15 NASDAQ National Market securities in December, 1982, and all The NASDAQ National Market securities in February, 1986.

as trading inside the spread (Lee, 1993). As quoted spreads represent a natural upper bound for effective spreads, I use the minimum between the effective spread estimate in each month and the average CRSP quoted spread to further improve the estimation accuracy. The final dataset consists of 14,347 stocks in the sample period 1982–2021, for a total of 1,304,784 stock-month observations. For each stock-month, I compute the ratio between the spread and the volatility per trade and Figure 3.3 reports the distribution of the ratio. Most of the values fall in the region between 2 and 3 and the distribution is peaked around the value 2.5, in close agreement with the model's prediction.

[Insert Figure 3.3 about here.]

To test the linear relation, I proceed as follows. First, I group observations into 100 bins based on the value of the volatility per trade.[7] Then, for each bin, I compute the average value of the spread and of the volatility per trade. Figure 3.4 plots the spread as a function of the volatility per trade on a double-logarithmic scale. The linear relation is satisfied with high accuracy across several orders of magnitude, for volatilities and spreads ranging from 0.01% to 10%. Regressing the logarithm of the spread on the logarithm of the volatility per trade yields estimates that support the linear relation with a coefficient of proportionality of about 2.4.

$$\log(S_t) = \underset{[0.01]}{0.97} \log\left(\frac{\sigma_t}{\sqrt{N_t}}\right) + \underset{[0.05]}{0.86} \quad \longrightarrow \quad S_t = e^{0.86}\left(\frac{\sigma_t}{\sqrt{N_t}}\right)^{0.97} \approx \frac{2.4\sigma_t}{\sqrt{N_t}} \tag{3.24}$$

[Insert Figure 3.4 about here.]

3.3.2 Estimating Trading Frequency from Price Data

This is the first paper that documents the possibility to estimate the trading frequency from price data. To empirically validate Equation (3.20), I proceed as follows. First, I compute the standard deviation of daily returns for each stock within each month. Second, I use the spread estimates introduced in Section 3.3.1. Third, I estimate the average number of daily trades for each stock-month by computing the ratio between volatility and spread as in Equation (3.20), where I set the coefficient of proportionality $2/(\lambda_t f_t)$ equal to 2.5 as obtained in Section 3.2.6. Zero-spread estimates, which would lead to an undefined number of trades, are dropped. Finally, I compare these estimates with the CRSP benchmark for securities listed in the NASDAQ Stock Market. For each stock-month, the benchmark is the average number of trades made on each day as provided by CRSP. The dataset consists of 1,102,254 stock-month observations in the sample period 1982–2021. I find that the overall correlation between the estimates and the corresponding CRSP benchmark is 89%.[8]

Figure 3.5 reports the time-series of the average number of trades per stock-day for the NASDAQ Stock Market. According to CRSP, stocks were traded on average 10 times

[7] The results are fully consistent by changing the number of bins or grouping by spread.

[8] As the number of trades varies widely in time and across stocks, the correlation is computed on its logarithm $\mathbb{C}\text{or}[\log(\hat{N}_t), \log(N_t)]$.

per day in the 1990s. This number has dramatically increased up to a peak of 10,000 trades per day in 2021. The estimates obtained from Equation (3.20) closely follow this trend. The figure also reports the number of trades obtained with Equation (3.21), which is based on bid-ask spreads computed via the Roll (1984) estimator. This series is unable to reproduce the trend of the CRSP benchmark. The reason is that the Roll (1984) estimator is affected by a large upward bias in the estimation of bid-ask spreads in small samples (Ardia et al., 2022), which induces a large downward bias in the calculation of the number of trades. These results confirm that Equation (3.20) holds with remarkable accuracy, as long as accurate estimates are available for the effective bid-ask spreads.

[Insert Figure 3.5 about here.]

To show the potential use of this measure, I estimate the number of trades also for NYSE and AMEX, for which such information is missing in CRSP. Figure 3.6, Panel A, reports the results for the full sample since 1927. I find that the trading frequency has remained rather stable in the last century, when the average number of trades per stock-day ranges between 10 and 100 for all markets. After the introduction of electronic trading in early 2000s, the number of trades has significantly increased. In the last decade, stocks in the AMEX market trade, on average, between 100 and 1,000 times per day. NASDAQ stocks trades between 1,000 and 10,000 times per day, and NYSE stocks are in the range between 10,000 and 100,000 daily trades. Panel B reports the average trade size, adjusted for inflation.[9] Here, I find that NYSE was used to execute the largest transactions. The average trade size increased from a minimum of 10,000 dollars in the early 1930s up to 1 million in early 2000s. After the introduction of electronic trading, the trade size has reduced and has become more homogeneous across market venues. In the last decade, I find that all markets execute average trades of 10,000 dollars.

[Insert Figure 3.6 about here.]

These estimates are in line with Brennan & Subrahmanyam (1998) who report an average trade size of 36,307 in 1988 for a subsample of 1,496 NYSE stocks (Brennan & Subrahmanyam, 1998, Table 1). Adjusting for inflation, that value corresponds to a trade size of 84,111 in real dollars as of December 2021, which is on the same order of magnitude of 100,000 real dollars in Figure 3.6.[10] The median size is also close. I obtain an estimate of 70,391 in real dollars, in good agreement with their reported value of 24,345 nominal dollars which is about 56,400 real dollars. Brennan & Subrahmanyam (1998) further report the average number of transactions. In 1988, each stock was traded on average 14,621 times during the year. This corresponds approximately to $14,621/252 = 58$ trades per day, in close agreement with Figure 3.6.

[9] The trade size is adjusted for inflation using the Consumer Price Index (CPI) provided by the Federal Reserve Bank of St. Louis (`https://fred.stlouisfed.org/series/CPIAUCNS`), so that 1 dollar unit corresponds to 1 dollar in real terms as of December 2021.

[10] The index as of December 2021 is $\text{CPI}_{2021} = 278$. The index as of December 1988 is $\text{CPI}_{1988} = 120$. Thus, adjusting for inflation gives $36307 \times \text{CPI}_{2021} \,/\, \text{CPI}_{1988} = 84111$.

3.4 Conclusion

This paper presents three propositions that describe the peak and permanent components of price impact and the price formation process. The predictions derived from the model are many, and they may open up new avenues for research. Firstly, the novel predictions proposed in this paper should be subject to thorough scrutiny and empirical evaluation across various markets and time periods. This extensive empirical work would allow determining the accuracy of the predictions under different circumstances. Secondly, the predictions are derived under the general assumption that trading occurs around the midpoint, resulting in the linearization of the order-book density. The prediction accuracy may be increased by solving Propositions I and II for higher-order approximations of the density or specific functional forms to account for the impact of orders executed deeper in the book. Another important point is to understand how the model formally generalizes from single to aggregate transactions (metaorders). Additionally, future research could extend the model by introducing trading frictions, such as commission fees or finite tick sizes, to examine how the predictions differ from those of a frictionless market.

The theory has broad implications that are detailed in the paper. Here, I focus the discussion on the main message: Trading forms prices. Information, behaviors, or any other reason that motivates people to trade are incorporated into prices through the impact of trades. But how to reconcile this view with the idea that the asset price is the discounted value of future payoffs? This idea states that the price is latent because of the limits of human knowledge that prevent us from observing the exact state of the economy. However, a superhuman agent with unbound knowledge may, in principle, determine the latent price, and observed prices will eventually converge to that price. In this view, trading does not form prices but instead discovers the latent price. Or, in words of Hasbrouck (2007), "orders do not impact prices. It is more accurate to say that orders forecast prices."

The theory of a latent price is consistent with the deterministic view of our universe that emerged in classical physics during the 18th century. According to determinism, exact knowledge of the state of the universe at a single instant of time would make it possible to reconstruct the past and predict the future from the laws of classical mechanics.[11] From this perspective, it is reasonable to assume that prices are predetermined and trading discovers them instead of forming them.

However, the view of the physical world has significantly evolved, and determinism has met several inconsistencies. In particular, it is incompatible with quantum mechanics, one of the most profound scientific developments of the 20th century (Kleppner & Jackiw, 2000). According to quantum mechanics, nature is intrinsically random. The ideas formalized by Bell (1964) have led to practical experiments to distinguish between fundamental randomness and randomness induced by not having complete information

[11]This articulation is known as the Laplace's demon. From the English translation of Laplace (2012): "We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."

about hidden variables. To date, experimental evidence has established that the hypothesis of hidden variables is inconsistent with the behavior of physical systems, and the determinism of classical physics is incapable of reproducing all the predictions of quantum physics.[12] From this perspective, it is impossible to know the exact state of the economy, and latent prices do not exist. Trading form prices because there is no price to discover.

The proposed theory aligns more with the contemporary understanding of the physical world. There is no latent price waiting to be discovered, and prices are formed in a purely mechanical manner through trading. Any motive that drives people to trade participates in the formation of prices, and financial markets reflect it. For example, investors who base their trades on fundamental analysis contribute to the maintenance of prices linked to the asset's intrinsic value (Fama, 1970). On the other hand, investors who trade based on behavioral biases introduce irrationality into the market (Shiller, 2003), while those trading based on liquidity needs factor in the cost of immediacy (Amihud, 2002). Investors who base their trades on socially responsible investing incorporate ethical considerations, and those who trade based on climate investing prioritize sustainability (Pástor et al., 2021). This approach complements the basic idea that the asset price reflects the discounted value of future payoffs, as that idea can be viewed as a particular case of the theory presented here where all investors base their trades on future payoffs discounted by either objective or subjective rates. This perspective suggests a new research path in which the age-old question of what drives prices is replaced with what motivates people to trade in the first place.

[12]See Nobel Prize Outreach AB 2023. Sat. 25 Feb 2023. https://www.nobelprize.org/prizes/physics/2022/press-release/

Bibliography

Abdi, F., & Ranaldo, A. (2017). A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies*, *30*(12), 4437–4480.

Almgren, R., Thum, C., Hauptmann, E., & Li, H. (2005). Direct estimation of equity market impact. *Risk*, *18*(7), 58–62.

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

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Labys, P. (2000). Great realizations. *Risk*, *13*, 105–108.

Andersen, T. G., Bondarenko, O., Kyle, A. S., & Obizhaeva, A. A. (2018). Intraday trading invariance in the e-mini s&p 500 futures market. *Anna A., Intraday Trading Invariance in the E-Mini S&P*, *500*.

Ardia, D., Guidotti, E., & Kroencke, T. A. (2022). Efficient estimation of bid-ask spreads from open, high, low, and close prices. SSRN. Working paper 3892335.

Bacidorea, J., Ross, K., & Sofianosa, G. (2003). Quantifying market order execution quality at the New York stock exchange. *Journal of Financial Markets*, *6*, 281–307.

Bacry, E., Iuga, A., Lasnier, M., & Lehalle, C.-A. (2015). Market impacts and the life cycle of investors orders. *Market Microstructure and Liquidity*, *1*(02), 1550009.

Bae, K.-h., Kyle, A. S., Lee, E. J., & Obizhaeva, A. A. (2016). Invariance of buy-sell switching points. *Invariance of Buy-Sell Switching Points (October 26, 2016). Robert H. Smith School Research Paper No. RHS*, *2730770*.

Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, *1*(3), 195.

Benzaquen, M., Donier, J., & Bouchaud, J.-P. (2016). Unravelling the trading invariance hypothesis. *Market Microstructure and Liquidity*, *2*(03n04), 1650009.

Bershova, N., & Rakhlin, D. (2013). The non-linear market impact of large trades: Evidence from buy-side order flow. *Quantitative Finance*, *13*(11), 1759–1778.

Bessembinder, H., & Kaufman, H. M. (1997). A comparison of trade execution costs for NYSE and NASDAQ-listed stocks. *Journal of Financial and Quantitative Analysis*, *32*(3), 287–310.

Bessembinder, H., & Seguin, P. J. (1993). Price volatility, trading volume, and market depth: Evidence from futures markets. *Journal of Financial and Quantitative Analysis*, *28*(1), 21–39.

Blanchard, O. J., & Watson, M. W. (1982). Bubbles, rational expectations and financial markets. National Bureau of Economic Research. Working paper 945.

Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Brennan, M. J., & Subrahmanyam, A. (1998). The determinants of average trade size. *The Journal of Business*, *71*(1), 1–25.

Brokmann, X., Serie, E., Kockelkoren, J., & Bouchaud, J.-P. (2015). Slow decay of impact in equity markets. *Market Microstructure and Liquidity*, *1*(02), 1550007.

Bucci, F., Benzaquen, M., Lillo, F., & Bouchaud, J.-P. (2018). Slow decay of impact in equity markets: insights from the ancerno database. *Market Microstructure and Liquidity*, *4*(03n04), 1950006.

Bucci, F., Lillo, F., Bouchaud, J.-P., & Benzaquen, M. (2020). Are trading invariants really invariant? Trading costs matter. *Quantitative Finance*, *20*(7), 1059–1068.

Chan, K., & Fong, W.-M. (2000). Trade size, order imbalance, and the volatility–volume relation. *Journal of Financial Economics*, *57*(2), 247–273.

Chordia, T., Roll, R., & Subrahmanyam, A. (2005). Evidence on the speed of convergence to market efficiency. *Journal of Financial Economics*, *76*(2), 271–292.

Chordia, T., Roll, R., & Subrahmanyam, A. (2008). Liquidity and market efficiency. *Journal of Financial Economics*, *87*(2), 249–268.

Corwin, S. A., & Schultz, P. (2012). A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance*, *67*(2), 719–760.

Cujean, J., & Hasler, M. (2017). Why does return predictability concentrate in bad times? *Journal of Finance*, *72*(6), 2717–2758.

Donier, J., & Bonart, J. (2015). A million metaorder analysis of market impact on the bitcoin. *Market Microstructure and Liquidity*, *1*(02), 1550008.

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

Farmer, J. D., Gerig, A., Lillo, F., & Waelbroeck, H. (2013). How efficiency shapes market impact. *Quantitative Finance*, *13*(11), 1743–1758.

Gabaix, X., & Koijen, R. S. (2021). In search of the origins of financial fluctuations: The inelastic markets hypothesis. National Bureau of Economic Research. Working paper 28967.

Gomes, C., & Waelbroeck, H. (2015). Is market impact a measure of the information value of trades? Market response to liquidity vs. informed metaorders. *Quantitative Finance*, *15*(5), 773–793.

Grinold, R. C., & Kahn, R. N. (2000). *Active portfolio management*. McGraw Hill New York.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, *70*(3), 393–408.

Hasbrouck, J. (2007). *Empirical market microstructure*. Oxford University Press.

Holden, C. W., & Jacobsen, S. (2014). Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance*, *69*(4), 1747–1785.

Huang, R. D., & Stoll, H. R. (1994). Market microstructure and stock return predictions. *Review of Financial Studies*, *7*(1), 179–213.

Huberman, G., & Stanzl, W. (2004). Price manipulation and quasi-arbitrage. *Econometrica*, *72*(4), 1247–1275.

Jones, C. M., Kaul, G., & Lipson, M. L. (1994). Transactions, volume, and volatility. *Review of Financial Studies*, *7*(4), 631–651.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *Journal of Financial and Quantitative Analysis*, *22*(1), 109–126.

Kleppner, D., & Jackiw, R. (2000). One hundred years of quantum physics. *Science*, *289*(5481), 893–898.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, (pp. 1315–1335).

Kyle, A. S., & Obizhaeva, A. A. (2016). Market microstructure invariance: Empirical hypotheses. *Econometrica*, *84*(4), 1345–1404.

Kyle, A. S., Obizhaeva, A. A., & Tuzun, T. (2020). Microstructure invariance in U.S. stock market trades. *Journal of Financial Markets*, *49*, 100513.

Laplace, P.-S. (2012). *A philosophical essay on probabilities*. Courier Corporation.

Lee, C. M., & Ready, M. J. (1991). Inferring trade direction from intraday data. *Journal of Finance*, *46*(2), 733–746.

Lee, C. M. C. (1993). Market integration and price execution for NYSE-listed securities. *Journal of Finance*, *48*(3), 1009–1038.

Lillo, F., Farmer, J. D., & Mantegna, R. N. (2003). Master curve for price-impact function. *Nature*, *421*(6919), 129–130.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *Journal of Finance*, *71*(1), 5–32.

Milgrom, P., & Stokey, N. (1982). Information, trade and common knowledge. *Journal of Economic Theory*, *26*(1), 17–27.

Moro, E., Vicente, J., Moyano, L. G., Gerig, A., Farmer, J. D., Vaglica, G., Lillo, F., & Mantegna, R. N. (2009). Market impact and trading profile of hidden orders in stock markets. *Physical Review E*, *80*(6), 066102.

Nagel, S. (2012). Evaporating liquidity. *Review of Financial Studies*, *25*(7), 2005–2039.

Pástor, L., Stambaugh, R. F., & Taylor, L. A. (2021). Sustainable investing in equilibrium. *Journal of Financial Economics*, *142*(2), 550–571.

Petersen, M. A., & Fialkowski, D. (1994). Posted versus effective spreads. good prices or bad quotes? *Journal of Financial Economics*, *35*, 269–292.

Pohl, M., Ristig, A., Schachermayer, W., & Tangpi, L. (2020). Theoretical and empirical analysis of trading activity. *Mathematical Programming*, *181*, 405–434.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *Journal of Finance*, *39*(4), 1127–1139.

Said, E., Ayed, A. B. H., Husson, A., & Abergel, F. (2017). Market impact: A systematic study of limit orders. *Market Microstructure and Liquidity*, *3*(03n04), 1850008.

Said, E., Bel Hadj Ayed, A., Thillou, D., Rabeyrin, J.-J., & Abergel, F. (2021). Market impact: A systematic study of the high frequency options market. *Quantitative Finance*, *21*(1), 69–84.

Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *The American Economic Review*, *71*(3), 421–436.

Shiller, R. J. (2003). From efficient markets theory to behavioral finance. *Journal of Economic Perspectives*, *17*(1), 83–104.

Shiller, R. J. (2014). Speculative asset prices. *American Economic Review*, *104*(6), 1486–1517.

Tóth, B., Eisler, Z., & Bouchaud, J.-P. (2016). The square-root impace law also holds for option markets. *Wilmott*, *2016*(85), 70–73.

Tóth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., & Bouchaud, J.-P. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, *1*(2), 021006.

Wyart, M., Bouchaud, J.-P., Kockelkoren, J., Potters, M., & Vettorazzo, M. (2008). Relation between bid–ask spread, impact and volatility in order-driven markets. *Quantitative Finance*, *8*(1), 41–57.

Zarinelli, E., Treccani, M., Farmer, J. D., & Lillo, F. (2015). Beyond the square root: Evidence for logarithmic dependence of market impact on size and participation rate. *Market Microstructure and Liquidity*, *1*(02), 1550004.

Zumbach, G. (2004). How trading activity scales with company size in the ftse 100. *Quantitative Finance*, *4*(4), 441–456.

Figures

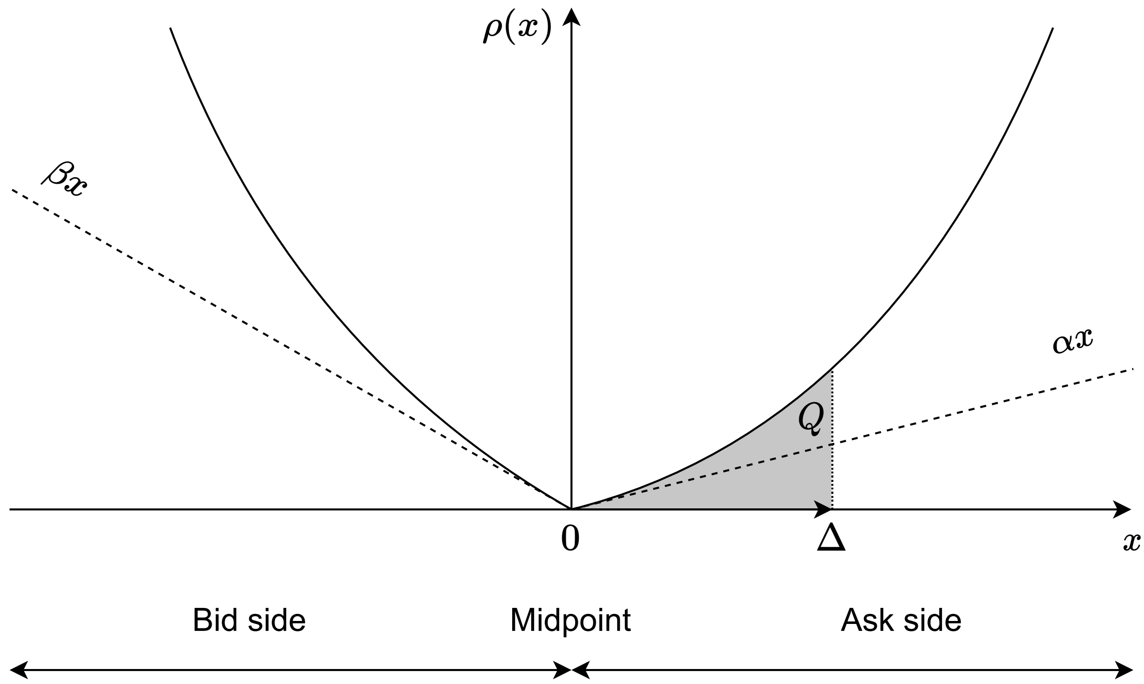


Figure 3.1: Density of the limit order book. The point $x = 0$ is the midpoint. The region $x > 0$ is the ask side and $x < 0$ is the bid side of the book. The coefficients α and β are the slopes of the lines tangent to the density in the midpoint. An order of size Q moves prices by a quantity Δ.

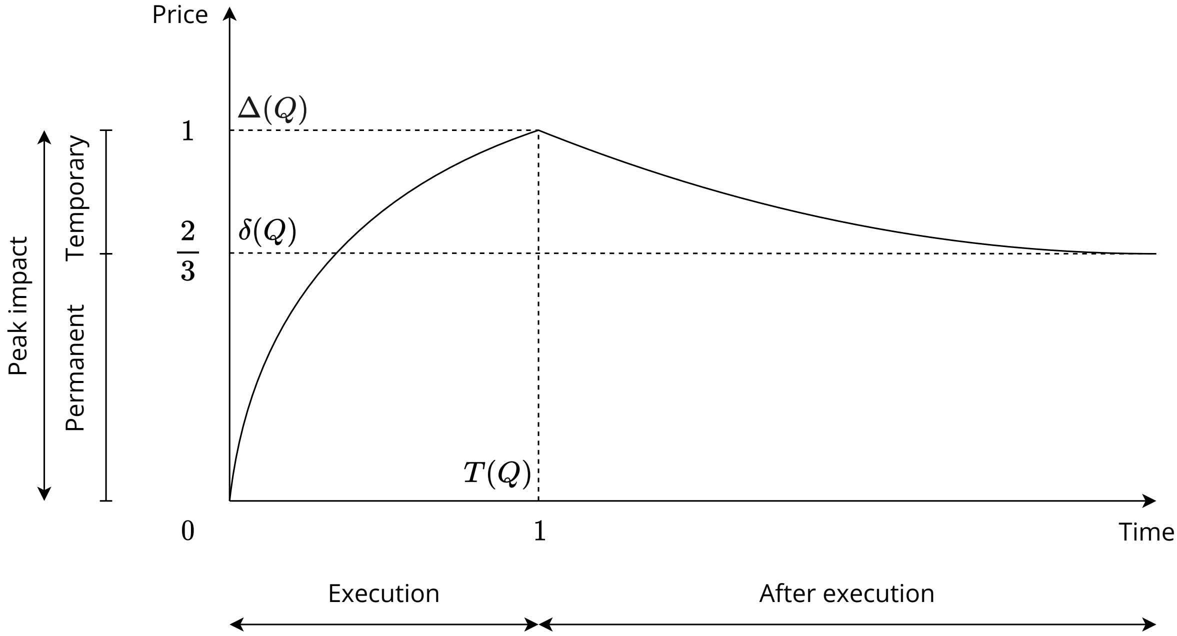


Figure 3.2: Permanent and temporary components of price impact. An order of size Q moves prices by a peak impact Δ. After execution, the peak impact reduces to the permanent impact δ. The peak impact increases as the square root of traded volume during execution, and the permanent impact is 2/3 of peak impact.

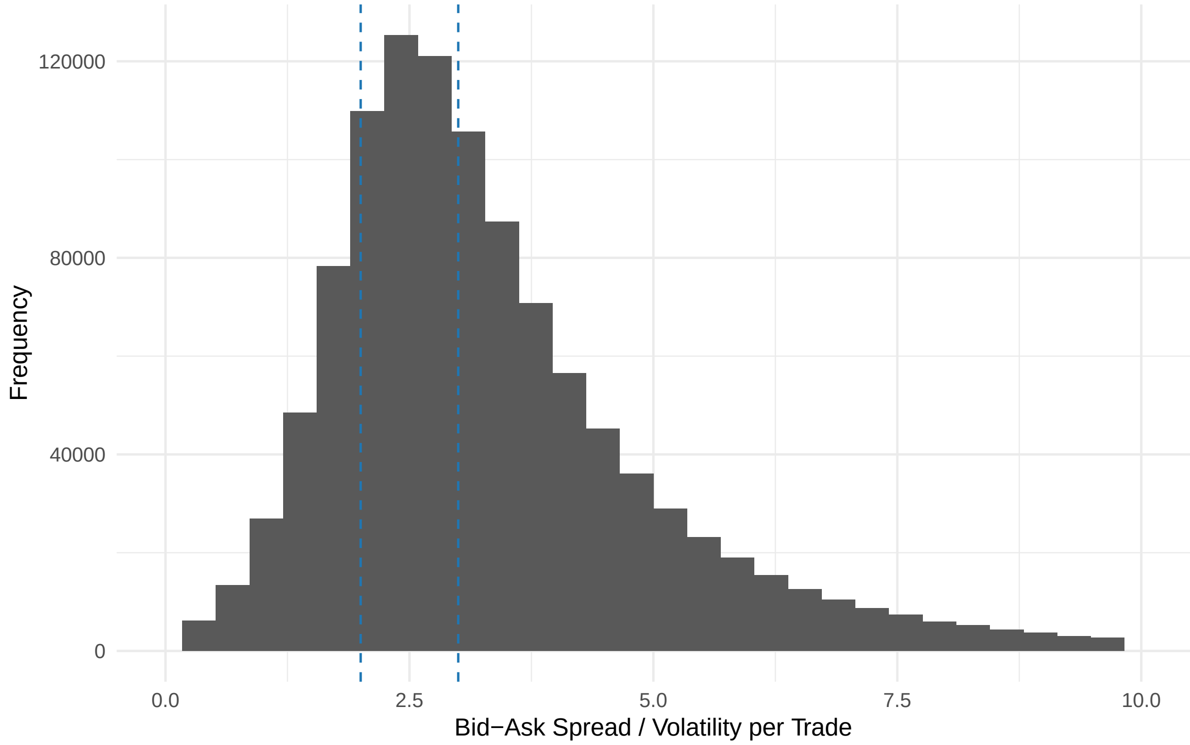


Figure 3.3: Distribution of the ratio between bid-ask spread and volatility per trade for NASDAQ stocks in the sample period 1982–2021. For each month, the bid-ask spread is calculated as the minimum between the average end-of-day quoted spread and the spread computed with the estimator by Ardia et al. (2022). Zero-spread estimates are dropped. The volatility per trade is calculated as the standard deviation of daily returns in each month divided by the square root of the average number of daily trades within the month.

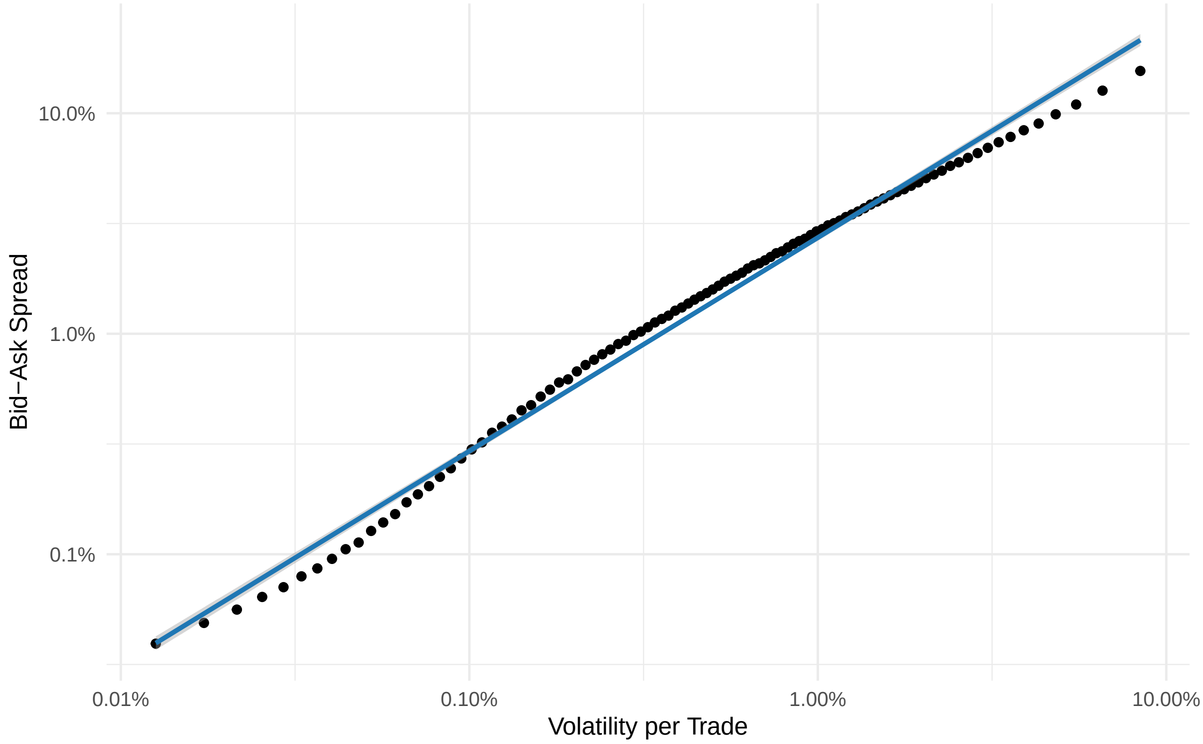


Figure 3.4: Relation between bid-ask spreads and volatility per trade for NASDAQ stocks in the sample period 1982–2021. For each month, the bid-ask spread is calculated as the minimum between the average end-of-day quoted spread and the spread computed with the estimator by Ardia et al. (2022). The volatility per trade is calculated as the standard deviation of daily returns in each month divided by the square root of the average number of daily trades within the month. Observations are grouped into 100 bins based on the value of the volatility per trade. For each bin, the figure reports the average value of the spread (y-axis) as a function of the average volatility per trade (x-axis) on a double-logarithmic scale. Averages are trimmed at the 1% level.

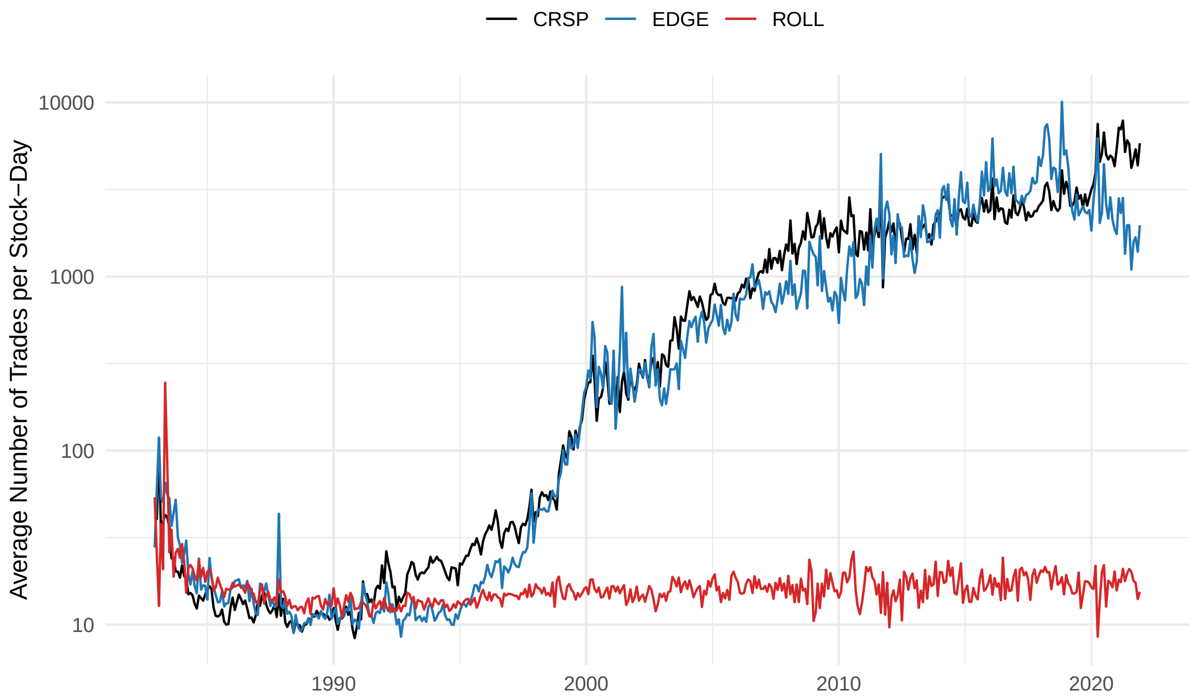
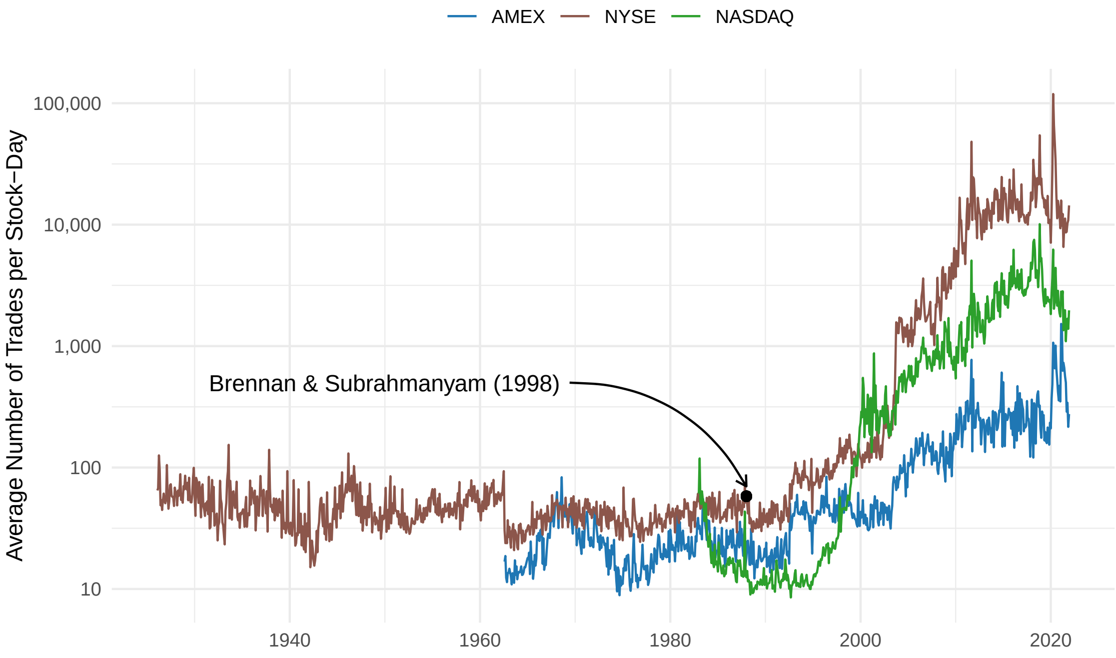
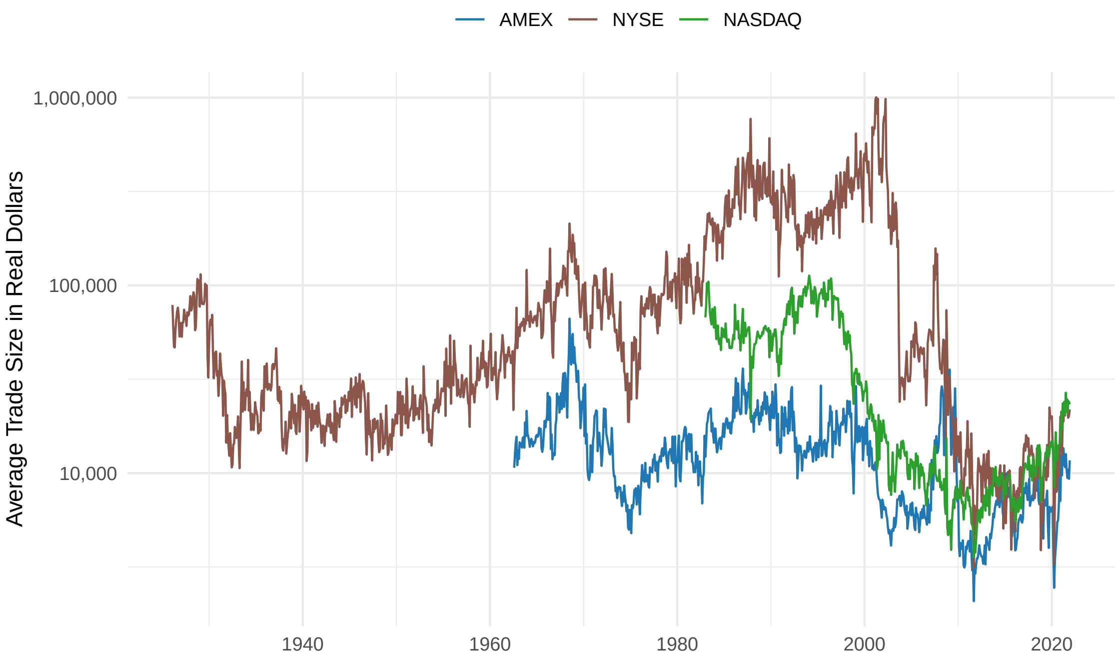


Figure 3.5: Time series of the average number of trades per stock-day in the NASDAQ Stock Market. Averages are trimmed at the 1% level. CRSP is the benchmark computed from the number of trades provided by the Center for Research in Security Prices. EDGE are the estimates obtained with Equation (3.20) as described in Section 3.3.2. ROLL are the estimates obtained with the Roll (1984) estimator as in Equation (3.21).



(a) Time-series of the average number of trades per stock-day in the AMEX, NYSE, and NASDAQ markets. Averages are trimmed at the 1% level. The estimates are obtained with Equation (3.20) as described in Section 3.3.2.



(b) Time series of the average trade size in the AMEX, NYSE, and NASDAQ markets. Averages are trimmed at the 1% level. For each stock-month, the trade size is computed dividing the average daily volume in dollars by the number of trades per day. The trade size is adjusted for inflation using the consumer price index provided by the Federal Reserve Bank of St. Louis (`https://fred.stlouisfed.org/series/CPIAUCNS`), so that 1 dollar unit corresponds to 1 dollar in real terms as of December 2021.

Figure 3.6: Number of trades and trade size for AMEX, NYSE, and NASDAQ.

3.A Appendix

3.A.1 Proofs

Proof of Equation (3.6)

Proof. Solve Equation (3.1) with the density in Equation (3.4).

$$Q = \int_0^\Delta cxdx = \frac{cx^2}{2}\Big|_0^\Delta = \frac{c\Delta^2}{2} \quad \text{with} \quad c = \begin{cases} \alpha & \text{if} \quad \Delta > 0 \\ \beta & \text{if} \quad \Delta < 0 \end{cases} \tag{3.25}$$

Thus, $\Delta = +\sqrt{2Q/\alpha}$ for buys and $\Delta = -\sqrt{2Q/\beta}$ for sells. □

Proof of Equation (3.7)

Proof. Solve Equation (3.2) with the density in Equation (3.4). For buys:

$$\int_0^\Delta (x-\delta)\alpha x dx = \alpha\Big(\frac{x^3}{3} - \frac{\delta x^2}{2}\Big)\Big|_0^\Delta = \alpha\Delta^2\Big(\frac{\Delta}{3} - \frac{\delta}{2}\Big) = 0\,. \tag{3.26}$$

Thus, $\delta = 2\Delta/3$. The same holds for sells by replacing α with β. □

Proof of Equation (3.9)

Proof. Equation (3.9) is obtained by substituting P_{j+n} and P_j with the corresponding expressions in Equation (3.3), and by rearranging the terms in the summations:

$$\begin{aligned} P_{j+n} - P_j &= \Big(P_0 + \sum_{i=1}^{j+n} \Delta_i(j+n-i)\Big) - \Big(P_0 + \sum_{i=1}^{j} \Delta_i(j-i)\Big) \\ &= \sum_{i=1}^{j} \Delta_i(j+n-i) - \Delta_i(j-i) + \sum_{i=j+1}^{j+n} \Delta_i(j+n-i)\,. \end{aligned} \tag{3.27}$$

□

Proof of Equation (3.11)

Proof. Let $g_j(n) = \mathbb{E}_j[w_j(n)]$ denote the expectation of $w_j(n)$ in Equation (3.9). The expectation generally depends on the impact Δ_j so that $g_j(n) = g_j(n, \Delta_j)$. A first-order expansion gives:

$$g_j(n) = g_j(n,0) + g_j'(n,0)\Delta_j\,, \tag{3.28}$$

where $g_j'(n,0)$ is the partial derivative of $g_j(n,\Delta_j)$ with respect to Δ_j evaluated in $\Delta_j = 0$. The component $\mu_j(n)$ in Equation (3.9) also depends on the impact Δ_j so that $\mu_j(n) = \mu_j(n, \Delta_j)$ and it can be rewritten as:

$$\mu_j(n) = \mu_j(n,0) + \Delta_j(n) - \Delta_j(0) = \mu_j(n,0) - \big(1 - f(n)\big)\Delta_j\,, \tag{3.29}$$

where $f(n)$ is defined in Equation (3.8). The expected return is thus given by:

$$\mathbb{E}_j[P_{j+n} - P_j] = \mu_j(n) + g_j(n) = \mu_j(n,0) + g_j(n,0) - \big(1 - f(n) - g'_j(n)\big)\Delta_j\,, \tag{3.30}$$

where $g'_j(n) = g'_j(n,0)$ for notational convenience. Conditioning on $Q_j = 0$, Equation (3.6) implies that $\Delta_j = 0$ and thus:

$$\mathbb{E}_j[P_{j+n} - P_j \mid Q_j = 0] = \mu_j(n,0) + g_j(n,0)\,. \tag{3.31}$$

Substituting Equation (3.31) in (3.30) gives Equation (3.11). □

Proof of Equation (3.12)

Proof. Define $a_j = \mathbb{E}_j[P_{j+n} - P_j \mid Q_j = 0]$ and $b_j = 1 - f(n) - g'_j(n)$ for notational convenience. The investor buys ($Q_j > 0$) when the expected return is positive ($a_j > 0$), and sells ($Q_j < 0$) when the expected return is negative ($a_j < 0$). For buys, Equation (3.11) implies that:

$$\mathbb{E}_j[P_{j+n} - P_j] = a_j - \Delta_j b_j = a_j - b_j\sqrt{\frac{2Q_j}{\alpha_j}}\,. \tag{3.32}$$

where Δ_j is replaced with Equation (3.6). The profits in Equation (3.10) are:

$$\pi(Q_j) = Q_j\mathbb{E}_j[P_{j+n} - P_j] = a_jQ_j - b_j\sqrt{\frac{2Q_j^3}{\alpha_j}}\,. \tag{3.33}$$

In what follows, the index j is redundant and omitted for readability. The stationary point is:

$$\pi' = a - 3b\sqrt{\frac{Q}{2\alpha}} = 0 \quad\longrightarrow\quad Q = \frac{2a^2\alpha}{9b^2}\,, \tag{3.34}$$

where π' is the derivative of π with respect to Q. The second-order derivative is positive for $b < 0$ and negative for $b > 0$:

$$\pi'' = \frac{-3b}{\sqrt{8\alpha Q}}\,. \tag{3.35}$$

Thus, the result in Equation (3.34) is profit-maximising for $b > 0$ but profit-minimizing for $b < 0$. In the latter case, the maximum profit is unbound and achieved for $Q \to \infty$. In the first case, substituting the expression for Q from Equation (3.34) into Equation (3.32) gives:

$$\mathbb{E}_j[P_{j+n} - P_j] = \frac{a}{3}\,. \tag{3.36}$$

The same holds for sells by using the corresponding impact from Equation (3.6) in Equation (3.32) and repeating the calculations. These results are summarized in Equation (3.12). □

Proof of Equation (3.14)

Proof. In a time period t, the probability of trades to be buyer-initiated is $p_t^{(buy)}$ and the probability to be seller-initiated is $p_t^{(sell)}$. The order-flow imbalance is $\phi_t = p_t^{(buy)} - p_t^{(sell)}$. From Equation (3.8), the impact function of the i-th trade is:

$$\Delta_i(n) = \begin{cases} +f(n)\sqrt{\dfrac{2Q_i}{\alpha_i}} & \text{with probability } p_t^{(buy)} = \dfrac{1+\phi_t}{2} \\ -f(n)\sqrt{\dfrac{2Q_i}{\beta_i}} & \text{with probability } p_t^{(sell)} = \dfrac{1-\phi_t}{2} \end{cases} \tag{3.37}$$

The expectation in Equation (3.5) is:

$$\begin{aligned} 0 &= \mathbb{E}[\Delta_i(n)] \\ &= p_t^{(buy)}\mathbb{E}_{buy}\left[f(n)\sqrt{\frac{2Q_i}{\alpha_i}}\right] + p_t^{(sell)}\mathbb{E}_{sell}\left[-f(n)\sqrt{\frac{2Q_i}{\beta_i}}\right] \\ &\approx p_t^{(buy)} f(n)\sqrt{\frac{2\mathbb{E}_{buy}[Q_i]}{\mathbb{E}_{buy}[\alpha_i]}} - p_t^{(sell)} f(n)\sqrt{\frac{2\mathbb{E}_{sell}[Q_i]}{\mathbb{E}_{sell}[\beta_i]}}\,. \end{aligned} \tag{3.38}$$

where $\mathbb{E}_{buy}$ and $\mathbb{E}_{sell}$ denote the expectations conditional on buys and sells, respectively. Rearranging the terms and writing $p_t^{(buy)}$ and $p_t^{(sell)}$ in terms of the order-flow imbalance ϕ_t gives:

$$\frac{1-\phi_t}{1+\phi_t} = \sqrt{\frac{\beta_t Q_t^{(buy)}}{\alpha_t Q_t^{(sell)}}}\,, \tag{3.39}$$

where $Q_t^{(buy)} = \mathbb{E}_{buy}[Q_i]$ is the average trade size of buys, $Q_t^{(sell)} = \mathbb{E}_{sell}[Q_i]$ is the average trade size of sells, $\alpha_t = \mathbb{E}_{buy}[\alpha_i]$ is the average α_i of buys, and $\beta_t = \mathbb{E}_{sell}[\beta_i]$ is the average β_i of sells. Taking the square on both sides and replacing α_t and β_t with γ_t and θ_t defined by Equation (3.13) gives Equation (3.14). □

Proof of Equation (3.15)

Proof. According to Equation (3.9), the return over the period is composed by the component $\mu_j(n)$, which is predictable by past impacts, and the component $w_j(n)$, which depends on future impacts. As $\mu_j(n)$ is easy to predict, this component shrinks to zero due to the action of skilled investors as described in Section 3.2.2. The variance of returns is thus the variance of $w_j(n)$, where $n = N_t$ is the number of trades in period t:

$$\begin{aligned}
\sigma_t^2 &= \mathbb{V}ar[P_{j+N_t} - P_j] \\
&= \mathbb{V}ar\left[\sum_{i=j+1}^{j+N_t} \Delta_i(j+N_t-i)\right] \\
&= \lambda_t^2 \times \sum_{i=j+1}^{j+N_t} \mathbb{V}ar\left[\Delta_i(j+N_t-i)\right] \\
&= \lambda_t^2 \times \sum_{i=j+1}^{j+N_t} \mathbb{E}\left[\Delta_i^2\right] f^2(j+N_t-i) \\
&= \lambda_t^2 \times \mathbb{E}\left[\Delta_i^2\right] \times \sum_{i=j+1}^{j+N_t} f^2(j+N_t-i) \\
&= \lambda_t^2 \times \mathbb{E}\left[\Delta_i^2\right] \times N_t \times f_t^2
\end{aligned} \tag{3.40}$$

where λ_t^2 and f_t^2 are given in Equation (3.16). The term $\mathbb{E}[\Delta_i^2]$ is computed using the peak impacts in Equation (3.6):

$$\begin{aligned}
\mathbb{E}[\Delta_i^2] &= p_t^{(buy)}\mathbb{E}_{buy}\left[\frac{2Q_i}{\alpha_i}\right] + p_t^{(sell)}\mathbb{E}_{sell}\left[\frac{2Q_i}{\beta_i}\right] \\
&\approx (1+\phi_t)\frac{Q_t^{(buy)}}{\alpha_t} + (1-\phi_t)\frac{Q_t^{(sell)}}{\beta_t} \, .
\end{aligned} \tag{3.41}$$

where ϕ_t is the order-flow imbalance, $Q_t^{(buy)} = \mathbb{E}_{buy}[Q_i]$ is the average trade size of buys, $Q_t^{(sell)} = \mathbb{E}_{sell}[Q_i]$ is the average trade size of sells, $\alpha_t = \mathbb{E}_{buy}[\alpha_i]$ is the average α_i of buys, and $\beta_t = \mathbb{E}_{sell}[\beta_i]$ is the average β_i of sells. The traded volume is the summation of all $|Q_i|$ in the period and it is equal to:

$$\begin{aligned}
V_t &= \sum_{i=1}^{N_t} |Q_i| \\
&= N_t^{(buy)}|Q_t^{(buy)}| + N_t^{(sell)}|Q_t^{(sell)}| \\
&= N_t\frac{1+\phi_t}{2}Q_t^{(buy)} - N_t\frac{1-\phi_t}{2}Q_t^{(sell)} \, .
\end{aligned} \tag{3.42}$$

Solving Equations (3.39) and (3.42) for $Q_t^{(buy)}$ and $Q_t^{(sell)}$ gives:

$$\begin{aligned}
Q_t^{(buy)} &= \frac{2V_t}{N_t}\frac{\alpha_t(1-\phi_t)}{\alpha_t(1-\phi_t)(1+\phi_t) - \beta_t(1+\phi_t)^2} \, , \\
Q_t^{(sell)} &= \frac{2V_t}{N_t}\frac{\beta_t(1+\phi_t)}{\alpha_t(1-\phi_t)^2 - \beta_t(1-\phi_t)(1+\phi_t)} \, .
\end{aligned} \tag{3.43}$$

Substituting Equation (3.43) in (3.41) and using Equation (3.13) gives:

$$\mathbb{E}[\Delta_i^2] = \frac{4V_t}{N_t}\frac{1}{\alpha_t(1-\phi_t) - \beta_t(1+\phi_t)} = \frac{2V_t}{N_t}\frac{1}{\gamma_t(1+\theta_t\phi_t)} \, . \tag{3.44}$$

Substituting Equation (3.44) in (3.40) gives (3.15).

□

Proof of Equation (3.17)

Proof. Compute γ_t from Equation (3.15):

$$\gamma_t = \frac{2V_t\lambda_t^2 f_t^2}{\sigma_t^2(1+\theta_t\phi_t)} \,. \tag{3.45}$$

Multiply and divide Equation (3.8) by $\sqrt{\gamma_t/\gamma_t}$:

$$\Delta_i^{(buy)}(n) = +f(n)\sqrt{\frac{2Q_i}{\alpha_i}\frac{\gamma_t}{\gamma_t}}\,, \qquad \Delta_i^{(sell)}(n) = -f(n)\sqrt{\frac{2Q_i}{\beta_i}\frac{\gamma_t}{\gamma_t}}\,. \tag{3.46}$$

Substituting γ_t at the denominator in Equation (3.46) with (3.45) gives (3.17). $\square$

Proof of Equation (3.19)

Proof. As $S_i = 2|\Delta_i|$, then $\Delta_i = D_iS_i/2$ where D_i is an indicator variable that equals $+1$ if the trade is a buy and -1 if the trade is a sell. The variance of peak impacts is:

$$\mathbb{E}[\Delta_i^2] = \mathbb{E}\left[\left(\frac{D_iS_i}{2}\right)^2\right] = \frac{\mathbb{E}[S_i^2]}{4} = \frac{S_t^2}{4}\,, \tag{3.47}$$

where S_t is the root-mean-square spread in period t. Substituting this expression in Equation (3.40) gives:

$$\sigma_t^2 = \frac{N_tS_t^2\lambda_t^2 f_t^2}{4}\,. \tag{3.48}$$

Solving for S_t gives Equation (3.19). $\square$

Proof of Equation (3.20)

Proof. Solving Equation (3.19) for N_t gives (3.20). $\square$

Proof of Equation (3.21)

Proof. The Roll (1984) estimator for the bid-ask spread is:

$$S_t = 2\sqrt{-\mathbb{C}\mathrm{ov}[R_t, R_{t-1}]} \tag{3.49}$$

where R_t is the return in period t and R_{t-1} is the return in the previous period. Substituting this expression in Equation (3.20) gives:

$$\frac{\mathbb{C}\mathrm{ov}[R_t, R_{t-1}]}{\sigma_t^2} = \frac{-1}{\lambda_t^2 f_t^2 N_t}\,. \tag{3.50}$$

Equation (3.21) is obtained by noting that $\sigma_t^2 = \mathbb{V}ar[R_t]$ and, thus, the left-hand side of Equation (3.50) is equal to $\mathbb{C}\mathrm{or}[R_t, R_{t-1}]$. $\square$

Proof of Equation (3.22)

Proof. From Equation (3.19):

$$\sigma_t = \frac{\lambda_t f_t S_t \sqrt{N_t}}{2}\,. \tag{3.51}$$

Substituting Equation (3.51) in Equation (3.15) gives:

$$V_t = \frac{S_t^2 N_t \gamma_t (1 + \theta_t \phi_t)}{8}\,. \tag{3.52}$$

From Equations (3.51)–(3.52):

$$\sigma_t V_t = \frac{\lambda_t f_t \gamma_t (1 + \theta_t \phi_t)}{16} S_t^3 N_t^{3/2}\,. \tag{3.53}$$

Rearranging the terms gives Equation (3.22). □