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

DAVID ARDIA, EMANUELE GUIDOTTI[*] and TIM A. KROENCKE [†]

October 18, 2021

Abstract

We propose a novel estimation procedure of bid-ask spreads from open, high, low, and close prices. Our estimator is asymptotically unbiased and optimally combines the full set of price data to minimize the estimation variance. When quote data are not available, our estimator generally delivers the most accurate estimates of effective bid-ask spreads numerically and empirically. The estimator is derived under permissive assumptions that allow for stylized facts typically observed in real market data, is easy to implement, and can be applied to liquid and illiquid market segments, both in low and high frequency.

[*]David Ardia is at Department of Decision Sciences and GERAD, HEC Montréal. Emanuele Guidotti and Tim A. Kroencke are at Institute of Financial Analysis, University of Neuchâtel. E-mail addresses: david.ardia@hec.ca (D. Ardia), emanuele.guidotti@unine.ch (E. Guidotti), tim.kroencke@unine.ch (T. A. Kroencke).

[†]Earlier versions of this paper were presented at the 27th Annual Meeting of the German Finance Association (DGF), and the 2021 International Risk Management Conference. We thank Dennis Umlandt (DGF discussant) and Angelo Ranaldo for their helpful comments. We acknowledge financial support by the Institute for Data Valorization (IVADO).

THE BID-ASK SPREAD is one of the predominant measures of liquidity in finance, with applications ranging from asset pricing (*e.g.*, Amihud and Mendelson, 1986; Korajczyka and Sadka, 2008) to corporate finance (*e.g.*, Barclay and Smith Jr., 1988; Fang, Tian, and Tice, 2014) and accounting research (*e.g.*, Dechow, Sloan, and Sweeney, 1996; Blankespoor, deHaan, and Marinovic, 2020). However, a direct computation of the effective bid-ask spread requires to match high-frequency trade and quote data (Holden and Jacobsen, 2014), which are typically not available for international markets, asset classes other than stocks, and time periods prior to 1993 (Corwin and Schultz, 2012; Abdi and Ranaldo, 2017). The size of the quoted bid-ask spread, a popular proxy for the effective spread, is even more fraught with measurement problems. The quoted spread has been shown to overestimate the effective spreads finally paid by traders by up to 100% (see, *e.g.*, Huang and Stoll, 1994; Petersen and Fialkowski, 1994; Bessembinder and Kaufman, 1997; Bacidorea, Ross, and Sofianosa, 2003), due to dealers offering a better price than the quotes, also known as trading inside the spread (Lee, 1993). Accordingly, most studies either limit the sample to the periods and markets of common data coverage or use liquidity proxies estimated from price data only (Hasbrouck, 2009).

Following the seminal work by Roll (1984), several approaches have been proposed to estimate the effective bid-ask spread by relying solely on readily available daily prices.[1] Among them, the estimators by Corwin and Schultz (2012) and Abdi and Ranaldo (2017) stand out, as they have been shown to generally deliver the most accurate estimates of effective spreads, both numerically and empirically (Corwin and Schultz, 2012; Holden and Jacobsen, 2014; Karnaukh, Ranaldo, and Söderlind, 2015; Abdi and Ranaldo, 2017; Johann and Theissen, 2017).

In this paper, we propose an *Efficient Discrete Generalized Estimator* (EDGE) of the

[1] Hasbrouck (2009) proposes a Gibbs estimation of the Roll model that is based on daily closing prices. Lesmond, Ogden, and Trzcinka (1999) introduce the LOT model that requires only the time series of daily security returns to endogenously estimate the effective transaction costs for any firm, exchange, or time period. Fong, Holden, and Trzcinka (2017) develop a new percent-cost proxy (FHT) which simplifies the existing LOT measure. Goyenko, Holden, and Trzcinka (2009) develop a proxy of the effective spread based on observable price clustering.

bid-ask spread that builds on –and improves– estimators based on transaction prices, in particular the influential contributions by Roll (1984), Corwin and Schultz (2012), and Abdi and Ranaldo (2017). Our contribution to the literature is twofold.

First, we develop a generalized methodology that allows us to derive bid-ask spread estimators from several combinations of Open, High, Low, and Close (OHLC) prices when trading is discrete. As two special cases, our methodology produces the estimators in Roll (1984) and Abdi and Ranaldo (2017) with a correction term for infrequent trading. Although the previous literature has focused on continuous-time models (*e.g.*, Geometric Brownian Motion), we show that this practice leads to a significant *downward bias* when trading is infrequent (*i.e.*, for illiquid assets).[2] Instead, our generalized estimators remain unbiased. This is an important property, as illiquid assets are expected to have the highest transaction costs. In particular, a systematic underestimation of transaction costs, where they are expected to be the largest, may raise identification concerns regarding previous empirical findings.

Second, we provide the optimal way to combine our estimators to minimize the estimation variance and obtain an efficient estimator (EDGE). By exploiting the full set of OHLC prices, our efficient estimator is, on average, twice as accurate as the best performing estimators available today.[3] The increased accuracy allows us to produce estimates closer to the (true but unobserved) effective spread. Moreover, this property helps mitigating another *upward bias* that has been recently investigated by Jahan-Parvar and Zikes (2019) and Tremacoldi-Rossi and Irwin (2019). Negative estimates are usually re-set to zero to guarantee non-negativity of the transaction costs estimates (Goyenko, Holden, and Trzcinka, 2009; Hasbrouck, 2009; Corwin and Schultz, 2012; Karnaukh, Ranaldo, and Söderlind, 2015; Abdi and Ranaldo, 2017) and this practice

[2] For instance, the Corwin and Schultz (2012) liquidity measure has been translated to the corporate bond market by Schestag, Schuster, and Uhrig-Homburg (2016). As most bonds are infrequently traded, Nieto (2018) shows that this practice can produce an important bias, even when bonds with high activity requirements are selected.

[3] Abdi and 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 and Schultz, 2012) prices independently. To the best of our knowledge, EDGE is the first estimator exploiting the full information set of OHLC price data.

leads, on average, to overstating the spread. By reducing the estimation variance, we find that EDGE naturally produces a smaller fraction of negative estimates compared to the alternative estimators. Our results show that the number of negative estimates is reduced by 10% for smaller spreads and up to 50% for larger spreads.

Another advantage of EDGE is that it can be applied at any frequency and can exploit high-frequency price data whenever available. While the variance component of an asset return is proportional to the return interval, the spread component is not (Corwin and 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. We show that EDGE can estimate intraday spreads from minute data, while the other estimators struggle as trading becomes infrequent at this time interval, and their downward bias dominates the spread estimate. This property allows to study bid-ask spreads in high-frequency (*e.g.*, Lee, Mucklow, and Ready, 1993) for markets that do not report bid and ask data (*e.g.*, Bryant and Haigh, 2004). Moreover, by relying solely on transaction prices, our estimator is not deceived by quote stuffing, that is, the practice where a large number of orders to buy or sell are placed and then canceled almost immediately in an attempt to manipulate the market through fake bidding (Egginton, Van Ness, and Van Ness, 2016).

We compare EDGE with Roll (1984), Corwin and Schultz (2012), and Abdi and Ranaldo (2017) in a comprehensive simulation study and with empirical data.

In our simulation experiments, we compare the correlation coefficient, Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), 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 ranging from 10 minutes to one day. We find that EDGE produces the highest time-series and cross-sectional correlation coefficients, lowest MAPE, and RMSE, 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 frequency.

Our empirical analysis compares EDGE with the benchmark estimators, with the end-of-day quoted spread available in the CRSP U.S. stock database for 1925–1942 and 1993–2020, and with the effective spread computed via the TAQ database for the period 1993–2020. We find that our simulation-based results carry over to the empirical data. EDGE is more correlated and considerably closer to the effective spread than all other estimators, both in time-series and cross-sectional studies. While EDGE remains unbiased, the benchmark estimators underestimate the effective spread for small and less liquid stocks. The difference is economically large: in the historical sample of 1925–1942, we find that EDGE is often about two times larger than the next best estimator from transaction data. Our results also confirm earlier research (Huang and Stoll, 1994) showing that quoted spreads considerably overestimate the effective spread.

In recent sample periods and for highly liquid stocks characterized by a tiny bid-ask spread (*e.g.*, large caps), we confirm that all the estimators are upward biased when using daily data to estimate monthly spreads (Jahan-Parvar and Zikes, 2019; Tremacoldi-Rossi and Irwin, 2019). When increasing the estimation window to one year, we find that EDGE is unbiased for a spread size as small as 0.10% under ideal conditions and for a spread size of 0.30% when overnight returns are included in simulations. If more accurate spread estimates are needed, we illustrate how EDGE can produce unbiased estimates of tiny spreads from intraday minute data.

EDGE admits a simple closed-form formula and is easy to calculate. To guarantee reproducibility of our work, we make available easy-to-use software for the R statistical environment (R Core Team, 2020) that implements all the estimators and all the results in this paper.[4]

[4]The code is available upon request from the authors.

1 Methodology

1.1 Setting

We rely on a set of assumptions that are comparable to earlier contributions in the literature (*e.g.*, Roll, 1984; Corwin and Schultz, 2012; Abdi and Ranaldo, 2017). More specifically, we assume a spread of $S\%$, which is constant over the estimation period. The observed prices P_t for buys are higher than the actual prices $\tilde{P}_t$ by half the spread, while observed prices for sells are lower than the actual value by half the spread. Buys and sells are equally likely. Finally, actual returns are uncorrelated.[5] We formalize our assumptions in the following model:

$$P_t = \tilde{P}_t(1 + S(B_t - 0.5)) \,, \tag{1}$$

where B_t is a Bernoulli random variable with probability of success 0.5. In logarithmic prices p_t, Equation (1) becomes:

$$p_t = \tilde{p}_t + Z_t \,, \tag{2}$$

where we define $Z_t = S(B_t - 0.5)$ for notational convenience.[6]

1.2 Derivation of Discrete Generalized Estimators

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

[5] The assumption of zero autocorrelation in returns is less restrictive than independence, which is popular among older contributions. In particular, we do not rule out the possibility that the squared returns are autocorrelated (*i.e.*, time-varying volatility and volatility clustering).

[6] As S is typically much smaller than 1, we approximate $\ln(1 + Z_t) \approx Z_t$ based on a first-order Taylor expansion.

Equation (2) and expand the covariance in the four terms:

$$\begin{aligned}\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] &= \mathbb{C}\text{ov}[\tilde{c}_t - \tilde{c}_{t-1}, \tilde{c}_{t-1} - \tilde{c}_{t-2}] \\ &\quad + \mathbb{C}\text{ov}[\tilde{c}_t - \tilde{c}_{t-1}, Z_{t-1} - Z_{t-2}] \\ &\quad + \mathbb{C}\text{ov}[Z_t - Z_{t-1}, \tilde{c}_{t-1} - \tilde{c}_{t-2}] \\ &\quad + \mathbb{C}\text{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}] \\ &= \mathbb{C}\text{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}]\,,\end{aligned} \tag{3}$$

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

$$\begin{aligned}\mathbb{C}\text{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}] &= \mathbb{C}\text{ov}[Z_t, Z_{t-1}] \\ &\quad + \mathbb{C}\text{ov}[Z_t, -Z_{t-2}] \\ &\quad + \mathbb{C}\text{ov}[-Z_{t-1}, Z_{t-1}] \\ &\quad + \mathbb{C}\text{ov}[-Z_{t-1}, -Z_{t-2}]\,.\end{aligned} \tag{4}$$

Since the random variables Z are independent for different trades, so far, the literature has assumed that the only non-vanishing term in Equation (4) is $\mathbb{C}\text{ov}[-Z_{t-1}, Z_{t-1}] = -\mathbb{V}[Z]$. However, we point out that this is valid only under the assumption of continuous trading. If trading is continuous, there is an infinite amount of trades taking place between time t and time s, and the random variable Z_t is independent from Z_s for any $s \neq t$. In practice, trades occur at discrete time and the same trade can originate different prices on the market. For instance, one trade can originate two subsequent closing prices if no trade has occurred in the second period. In these circumstances, the random variables Z are not independent at different times, as they are actually originated by the same trade. Assuming that the only non-vanishing term in Equation (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 allows one single trade to generate different prices, as it is

often the case of periods with few trades or no trades at all (*i.e.* illiquid assets).

When there is no trade, the market reports the previous closing price, and therefore Z_t and Z_{t-1} are the same random variable originated from the same trade. In this case $\mathbb{C}\text{ov}[Z_t, Z_{t-1}] = \mathbb{V}[Z]$. By the covariance decomposition formula, we have:

$$\mathbb{C}\text{ov}[Z_t, Z_{t-1}] = \mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}]\,, \tag{5}$$

where $\mathbb{P}[Z_t = Z_{t-1}]$ is the probability that the same trade generated both Z_t and Z_{t-1}. The same holds for:

$$\mathbb{C}\text{ov}[-Z_{t-1}, -Z_{t-2}] = \mathbb{V}[Z]\mathbb{P}[Z_{t-1} = Z_{t-2}]\,, \tag{6}$$

where $\mathbb{P}[Z_{t-1} = Z_{t-2}]$ is the probability that the same trade generated both Z_{t-1} and Z_{t-2}. Moreover, we have:

$$\mathbb{C}\text{ov}[Z_t, -Z_{t-2}] = -\mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}]\mathbb{P}[Z_{t-1} = Z_{t-2}]\,. \tag{7}$$

Assuming only close prices are available, we estimate the probability that two subsequent prices are generated by the same trade by counting the fraction of times, $\nu_{c=c}$, for which the closing prices over two subsequent time periods are equal:

$$\mathbb{P}[Z_t = Z_{t-1}] = \mathbb{P}[Z_{t-1} = Z_{t-2}] \hat{=} \nu_{c=c}\,. \tag{8}$$

We can now compute the covariances in Equations (3)–(4) by substituting the terms in Equations (5)–(8):

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

We obtain our final formula by computing the variance of Z in Appendix A.1:

$$\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] = -\frac{S^2}{4}(1 - \nu_{c=c})^2. \tag{9}$$

Equation (9) can be easily solved for the spread S:

$$S^2 = -\frac{4\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}]}{(1 - \nu_{c=c})^2}, \tag{10}$$

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 and 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. Here we take into account that one trade can originate contemporaneously the open, high, low, and close prices if it is the only trade in the period, and that one trade can originate both the high (low) and close (open) price if the closing (opening) trade is selected as the highest (lowest) price. The calculations are provided in Appendix A.2.

In Table 1, we summarize the various estimators. When $\nu_{c=c} = 0$ (*i.e.*, there is at least one trade observed for each period), 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 and Ranaldo (2017).[7]

The first innovation of the estimators displayed in Table 1 is that they account for infrequent trading. While it is well known that the estimators by Roll (1984), Corwin and Schultz (2012), and Abdi and Ranaldo (2017) lead to biased results when trading is infrequent, our generalized estimators remain unbiased in these situations. Therefore, they can be applied to a wide range of asset classes, in liquid and illiquid market segments, at low or high frequency.

[7]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.

The second innovation of the estimators shown in Table 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 about here.]

1.3 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 generalized estimator in Table 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 A.3.1). Second, we include prior knowledge in the optimal GMM weighting matrix to improve the efficiency in small samples (Appendix A.3.2). Third, we provide an estimator for $k = 4p(1-p)$ in Table 1, where p is the probability of the high price to be buyer initiated or, equivalently, the probability of the low price to be seller initiated (Appendix A.3.3).

Following the calculations in Appendix A.3, we derive our *Efficient Discrete Generalized Estimator* (EDGE) of the bid-ask spread:

$$S^2 = \frac{w_1 \mathbb{E}[X_1] + w_2 \mathbb{E}[X_2]}{w_1 w_2 (\nu_{o=h,l} + \nu_{c=h,l}) - 1/2}\,, \tag{11}$$

where $\nu_{o=h,l}$ and $\nu_{c=h,l}$ are given in Table 1; X_1, X_2 are the following vectors based on log-returns where we need to drop all the periods t with no trades such that $h_t = l_t = c_{t-1}$; and with the optimal weights provided below:

$$\begin{aligned} X_{1,t} &= (\eta_t - o_t)(o_t - c_{t-1}) + (o_t - c_{t-1})(c_{t-1} - \eta_{t-1})\,, \\ X_{2,t} &= (\eta_t - o_t)(o_t - \eta_{t-1}) + (\eta_t - c_{t-1})(c_{t-1} - \eta_{t-1})\,, \end{aligned} \tag{12}$$

$$w_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\,, \quad w_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad , \tag{13}$$

$$\sigma_1^2 = \mathbb{V}[X_1]\,, \quad \sigma_2^2 = \mathbb{V}[X_2]\,. \tag{14}$$

For estimation, the usual sample counterparts replace the expectations and variances, respectively.[8] We expect EDGE to provide superior performance compared to the estimators by Roll (1984), Corwin and Schultz (2012), and Abdi and 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.

1.3.1 Dealing with Negative Estimates

The estimate $\hat{S}^2$ in Equation (11) may become negative in finite samples. This is an issue as a negative squared spread is not mathematically nor economically meaningful. To guarantee non-negativity of the transaction costs estimate, we follow the common approach of truncating negative values (Goyenko, Holden, and Trzcinka, 2009; Hasbrouck, 2009; Karnaukh, Ranaldo, and Söderlind, 2015):

$$\hat{S} = \sqrt{\max\left\{0\,, \hat{S}^2\right\}}\,. \tag{15}$$

Jahan-Parvar and Zikes (2019) and Tremacoldi-Rossi and Irwin (2019) document that the practice of resetting negative estimates to zero leads to overstating the spread when estimating monthly spreads from daily data and where the true spread is 0.50% and smaller. In Section 2.2.3, 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. Another option to reduce the estimation variance and avoid negative estimates is to use high-frequency price data as

[8] To further improve the robustness of the estimates, the sample mean can be replaced with a robust estimator of the mean, such as the winsorized mean or the trimmed mean.

illustrated in Section 3.5.

1.3.2 Confidence Intervals

In this section, we derive the distribution of EDGE in Equation (11) to allow for proper hypothesis testing on the bid-ask spread. We start by noting that the estimator is written as an expectation so that by the Central Limit Theorem it will be asymptotically normally distributed.[9] As the asymptotic variance has to be estimated from the data, in small samples, the estimator is distributed according to a t-distribution with $n-1$ degrees of freedom, where n is the sample size:

$$\frac{\hat{S}^2 - S^2}{\sigma/\sqrt{n}} \sim t_{n-1}\,. \tag{16}$$

The sample standard deviation σ is given by:

$$\sigma = \frac{\sqrt{w_1^2\sigma_1^2 + w_2^2\sigma_2^2 + 2w_1w_2\sigma_{12}}}{1/2 - w_1w_2(\nu_{o=h,l} + \nu_{c=h,l})}\,, \tag{17}$$

where $\sigma_{12} = \mathbb{C}\text{ov}[X_1, X_2]$ and all the other terms are the same as in Equation (11).

As we have uncovered the distribution of S^2, we can now derive the confidence intervals of S. By exploiting the fact that the spread is positive, we know that the probability of the estimate S to be less than a given level s is equal to the probability of S^2 to be less than the squared level s^2. This equals the cumulative density function $\Phi_{n-1}(s^2)$ of the t-distribution in Equation (16) computed in s^2:

$$p = \mathbb{P}[S < s] = \mathbb{P}[S^2 < s^2] = \Phi_{n-1}(s^2)\,. \tag{18}$$

Equations (18) and (15) allow to obtain the quantiles associated with a probability level p by computing the inverse of the cumulative density function $\Phi_{n-1}^{-1}(p)$:

$$s_p^2 = \max\left\{0,\ \Phi_{n-1}^{-1}(p)\right\}\,, \quad s_p = \sqrt{s_p^2}\,. \tag{19}$$

[9] By Slutsky's theorem, we can treat as constants the weights w and the frequencies ν in Equation (11).

Finally, we use Equation (19) to obtain the critical values for S at a confidence level $1-\alpha$:

$$\left[\sqrt{\max\left\{0,\ \Phi_{n-1}^{-1}(\alpha/2)\right\}}, \sqrt{\max\left\{0,\ \Phi_{n-1}^{-1}(1-\alpha/2)\right\}}\right]. \tag{20}$$

1.3.3 Random Spread

When considering a random spread, the variance of Z becomes $\mathbb{E}[S^2]/4$ instead of $S^2/4$ as shown in Appendix A.1.1. By using $\mathbb{E}[S^2]/4$ instead of $S^2/4$ in Appendix A.2, it can be seen that all the equations in Table 1 hold more in general for random spreads by substituting $\mathbb{E}[S^2]$ to S^2 in the left-hand side of the equations. In other words, if the spread is random, then all our estimators are formally estimators for the mean squared spread. Moreover, in case the spread does not vary widely around its mean, we can approximate $\mathbb{E}[S^2] \approx \mathbb{E}[S]^2$ so that all the formulas in Table 1, and in particular Equation (11) become, at least approximately, estimators for the average (random) spread.

1.4 Using EDGE in Practice

We expect EDGE to be successfully applied out-of-the-box without performing any ad hoc adjustment or price manipulation. Our estimator should work well in practice as it is derived under permissive assumptions that allow for infrequent trading, time-varying volatility, fat tails, overnight jumps, and other effects observed in actual price data. The following section demonstrates the benefits of EDGE in a controlled environment using a comprehensive simulation experiment.

2 Simulation Study

In this section, we perform a Monte Carlo study to assess the accuracy and robustness of the EDGE in Equation (11). We compare the results with the estimators recently proposed by Corwin and Schultz (2012) and Abdi and Ranaldo (2017). Both papers define at least two versions of their estimators that handle negative spread estimates

in different ways. The first version sets negative estimates to zero and offers the most natural benchmark for our estimator. We refer to these versions as the CS (Corwin and Schultz, 2012) and AR (Abdi and Ranaldo, 2017) estimators, respectively.[10] 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 and Schultz, 2012) and AR2 (Abdi and Ranaldo, 2017) estimators, respectively.[11] The CS and CS2 estimators are adjusted for overnight returns as described in Corwin and Schultz (2012).

2.1 Setup

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

2.1.1 Low Frequency

We simulate 10,000 stock-months where each month consists of 21 days and where 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 second 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 at time t, then the previous Close at time $t-1$ is used as the Open, High, Low, and Close prices at time t.

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

2.1.2 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 at time t, then the previous Close at time $t-1$ is used as the Open, High, Low, and Close prices at time t.

2.2 Results

In Table 2, we report the results of the simulation study in the low-frequency setting of Section 2.1.1. Panel A shows the comparison where prices are assumed to be observed each minute, and overnight returns are not included. In these simulations, EDGE demonstrates to be on average twice more precise than AR or 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 0.34%, while the AR (CS) estimate is 0.71% (0.60%) with a standard deviation of 0.78% (0.50%). By estimating spreads as large as 1.21%, 1.44%, 1.45%, respectively, AR2, CS2, and Roll demonstrate to be not accurate for small spreads as already observed in Tremacoldi-Rossi and 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 estimator with better accuracy and lower bias in all scenarios. EDGE performs similar to the AR estimator for the smallest considered spread (0.50%), and is more accurate and more precise in all other scenarios.

[Insert Table 2 about here.]

In the Appendix (Table A.2), we replicate the experiment in Jahan-Parvar and 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 A.1, we extend the comparison to the high-frequency setting described in Section 2.1.2. 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 across the estimators from several perspectives.

2.2.1 Bias

In Figure 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 2.1.1 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 estimators. 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 numbers of trades per day, suggesting it works well in practice even in the case of illiquid assets or in high frequency when only a few trades are observed per minute. 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 2 and already documented in the original papers. In the Appendix (Figure A.4), we extend the comparison to the high-frequency setting described in Section 2.1.2, from which the same conclusions can be drawn.

[Insert Figure 1 about here.]

2.2.2 Variance

In Figure 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 2.1.1, 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) 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 A.5), we extend the comparison to the high-frequency setting described in Section 2.1.2, from which the same conclusions can be drawn.

[Insert Figure 2 about here.]

2.2.3 Negative Estimates

A major drawback of bid-ask spread estimators is the large number of negative estimates they produce for sample sizes typically encountered in financial studies. Although AR2 and CS2 try to mitigate this issue at the cost of introducing a large bias in the estimation of small spreads, this problem does not seem to be effectively improved with any adjustment proposed in the literature (Jahan-Parvar and Zikes, 2019; Tremacoldi-Rossi and Irwin, 2019).

In Figure 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 2.1.1, 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 times) even when using one year of daily data to compute the spread. AR and CS exhibit a similar behaviour, producing non-positive estimates between 20% and 30% of the times 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 3 about here.]

2.2.4 Confidence Intervals

In Figure 4, we assess the empirical performance of the confidence intervals provided in Equation (20). To this end, we run simulations consisting of 390 trades per day as described in Section 2.1.1, estimate the spread for each month, and compute the fraction of times in which the true spread is outside of the confidence intervals (false positive rate). We repeat this exercise for confidence levels $(1 - \alpha)$ ranging from zero to 100% and for several spread levels. We notice how the empirical false positive rate that we obtain is close to the exact theoretical value α for all the confidence levels and the different spreads. The result suggests that the distribution in Equation (16) and the corresponding confidence intervals in Equation (20) are reliable even in small samples with as little as 21 observations (monthly estimates from daily data).

[Insert Figure 4 about here.]

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 5, we simulate 10,000 stock-months with 21 days in each month under this setting. For each day, we use the previous year to estimate the spread. The estimates are benchmarked with the average 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, 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 volume 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 volume. The same problem would affect intraday spread estimates where the larger trading volume, usually observed around market opening and close, is likely to be artificially reflected on the spread. EDGE allows for a consistent comparison regardless of changes in the trading volume. 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 5 about here.]

In the Appendix (Table A.3), we report the correlation coefficient, Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), achieved by the estimators using a rolling window ranging from one year to one month. We also report the same metrics for simulations performed in high frequency, where we use an estimation window ranging from 10 minutes to one day. EDGE produces the highest correlation

coefficient, lowest MAPE, and RMSE, suggesting it is always the best choice regardless of the time interval and the evaluation metric used by the researcher, both in low and high frequency.

3 Empirical Results

In this section, we investigate how close we can estimate actual trading costs in the empirical data. To evaluate the performances of the estimators, we first need to define the ground truth, that is, the true value of the spread that serves as the benchmark for the evaluation. The simplest type of bid-ask spread is the quoted spread:

$$Q = 2\frac{Ask - Bid}{Ask + Bid}\,. \tag{21}$$

Quoted spreads are a popular measure, but they often overstate the true spreads finally paid by traders due to dealers offering a better price than the quotes, also known as trading inside the spread (Lee, 1993).

Effective spreads account for this issue by using trade prices. However, they are considerably more challenging to measure since one needs to match trades with quotes and account for reporting delays. Often the required data is not available. Effective spreads are defined as:

$$S = 2\frac{|P - M|}{M}\,, \quad M = \frac{Ask + Bid}{2}\,, \tag{22}$$

where P is the trade price and M is the midpoint computed from the bid and the ask prices. Following the literature, we use the effective spread to evaluate the performance of the various estimators that only requires commonly available OHLC price data.

3.1 Data Preparation

To compute the bid-ask spread estimates (*i.e.*, EDGE, AR, AR2, CS, CS2, Roll), we rely on the CRSP US Stock Database to access daily OHLC price data in the periods 1925–1962 and 1993–2020.[12] To compute the benchmark effective spread, we rely on the Trades and Quotes (TAQ) data available in 1993–2020. Effective spreads are obtained via the Wharton Research Data Services (WRDS) using Monthly TAQ for 1993–2003 and Daily TAQ from 2004 onward. The effective spreads are computed in the WRDS cloud according to the methodology described in Holden and Jacobsen (2014), which is also used in Abdi and Ranaldo (2017).[13] We match CRSP and TAQ data using CUSIP identifiers.[14] Our identification strategy allows us to match above 99.5% of the stocks in CRSP.

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 available.[15] Following Corwin and Schultz (2012) and Abdi and Ranaldo (2017), we select all NYSE, AMEX, and NASDAQ stocks with CRSP share codes of 10 or 11 (*i.e.*, U.S. common shares). No other data pre-processing is performed to maintain all the complexity of empirical data and especially of the highly illiquid stocks with only a few observations per month.

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. The monthly benchmark is computed as the average of the effective spreads in Equation (22) within the month. The minimal pre-processing allows us to cover a diverse and large sample of more than 1.6 million spread estimates for each estimator.

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

[13] The effective spreads computed via the methodology in Holden and Jacobsen (2014) are available to download from the WRDS Intraday Indicators.

[14] We reconstruct the time series of CUSIP for each KYPERMNO in CRSP. Then, we compute the time series of CUSIP for each stock in TAQ using the Monthy TAQ Master files for 1993–2009 and the Daily TAQ Master files in 2010–2020. Finally, we merge the datasets based on date and CUSIP.

[15] 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 drop these non-transaction-based observations.

In Table 3, we provide summary statistics of our empirical analysis based on 1993–2020 sample when CRSP and TAQ data are available. We report the mean, median, and standard deviation for the estimates and the effective spread benchmark. We notice how EDGE achieves the highest correlation (76.48%) with the benchmark, the smallest fraction of zero estimates (24.79%) and the lowest MAPE and RMSE.[16] The remainder of this section is dedicated to a deeper comparison across the estimators in a cross-sectional, time-series, and panel-data setting.

[Insert Table 3 about here.]

3.2 Cross-Sectional Correlation

Looking at cross-sectional correlations on a month-by-month basis allows us to evaluate the ability of the estimators in capturing 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 $\rho_t = \mathbb{C}\text{or}_i[S_{i,t}, \hat{S}_{i,t}]$. The month-by-month cross-sectional correlations for the various estimators are displayed in Figure 6. We see that the correlation between EDGE and the effective spread benchmark is consistently higher than the correlations achieved by AR, CS, or the Roll estimator throughout the whole period considered in the analysis.[17]

[Insert Figure 6 about here.]

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 in capturing 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.[18] Then, given the effective spread benchmark $S_{i,t}$ for stock i at time t

[16] 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 A.5.

[17] AR2 and CS2 perform similar to AR and CS and can be found in Table 4, Panel B.

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

and the corresponding estimate $\hat{S}_{i,t}$, we compute the time series correlation for decile d as $\rho_d = \mathbb{C}\text{or}_{i\in d,t}[S_{i,t}, \hat{S}_{i,t}]$. The time-series correlations for each decile obtained with the various estimators are displayed in Figure 7. We see that the correlation between EDGE and the effective spread benchmark is consistently higher than the correlations achieved by AR, CS, or the Roll estimator for all kinds of stocks.[19] The figure also shows a drop in the correlation associated with the very large stocks in the last decile, which are typically characterized by a tiny bid-ask spread. When more accurate estimates are needed for very large stocks, a researcher may consider using intraday data as illustrated in Section 3.5.

[Insert Figure 7 about here.]

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 and Schultz (2012) and Abdi and Ranaldo (2017). In addition, we extend the sample and include the more recent sub-period 2016–2020. For market capitalizations, we split the stocks in quintiles using the same procedure described in Section 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 $\rho_g = \mathbb{C}\text{or}_{(i,t)\in g}[S_{i,t}, \hat{S}_{i,t}]$.

The results are summarized in Table 4 for market venues (Panel A), time periods (Panel B), market capitalization (Panel C), and spread size (Panel D). One clear result emerges: EDGE outperforms all the alternative estimators in each market venue, sub-period, market capitalization, and spread size by consistently achieving the highest correlation with the TAQ effective spread benchmark.[20]

[19] AR2 and CS2 are similar to AR and CS and can be found in Table 4, Panel C.

[20] In Appendix A.5, we also provide the comparison on MAPE and RMSE and extend the estimation

Our results also highlight an overall tendency of all the estimators to perform poorer on more recent time periods, larger stocks, and smaller spread sizes (Jahan-Parvar and Zikes, 2019; Tremacoldi-Rossi and Irwin, 2019). Since the spread for certain stocks becomes smaller and smaller while the stock variance remains roughly the same, the spread becomes notoriously difficult to estimate from a given number of observations and the fraction of non-positive spread estimates increases.

[Insert Table 4 about here.]

3.5 Illustration of High Frequency Estimates

When the bid-ask spread is expected to be tiny (*i.e.*, below 0.50%), a researcher may consider increasing the estimation accuracy by using intraday price data to reduce the estimation variance and improve the spread estimates. 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 unbiased when trading becomes more and more infrequent. Indeed, we show earlier in the simulation experiment (Appendix Table A.1) that EDGE is expected to perform considerably better in such a scenario than other approaches.

To illustrate with empirical data, we show in Figure 8a the monthly spread estimates for GameStop Corp. (GME) in 2020, which is featured by a small effective spread of around 0.16% throughout the year. We find that the monthly estimates obtained from daily data vary wildly and tend to be significantly upward biased, while those obtained from hourly and minute data improve the accuracy of the estimates. In particular, as depicted in Figure 8b, the estimates obtained with minute data are sufficiently precise and allow the estimation of unbiased spreads from intraday prices even when using a daily estimation window.

[Insert Figure 8a and Figure 8b about here.]

window to one year. EDGE consistently achieves the highest correlation, lowest MAPE and RMSE for each sample size and evaluation metric.

4 Revisiting Historical Spread Estimates

To demonstrate the potential benefits of EDGE, we now turn to the analysis of historical trading costs using CRSP data since 1925. For each month, we construct 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-cap, mid-cap, and large-cap using the common 50th and 80th percentiles as breakpoints. Finally, we track the average spread of the three portfolios in 1925–1962 (CRSP sample) and 1993–2020 (CRSP-TAQ merged sample).

The results are reported in Figure 9 where small, mid, and large caps are shown in Panel A, B, and C, respectively. From the recent sample (CRSP-TAQ), we conclude that EDGE closely follows the effective spread whenever the transaction costs are not tiny. This is the case for small-cap stocks and all stocks before the year 2000. CS and AR tend to underestimate the transaction costs, particularly for small-cap stocks, mirroring the fact that these estimators are biased in the presence of low liquidity. Moreover, we find that the quoted spread overestimates the effective spread and does not constitute a reliable alternative.

In the arguably less liquid historical sample period 1925–1962, we find that the gap between EDGE and the alternative estimators further widens. The unbiased EDGE is by a factor of two larger than AR, and the difference is even more pronounced compared to CS. From this observation, we conclude that previous research based on low-frequency estimators has considerably underestimated transaction costs. Finally, we find that the quoted spread is considerably larger than EDGE in 1925–1962. Given our benchmark result from the recent sample, we conjecture that the quoted spread significantly overestimates the effective spread in the early sample. As TAQ data are not available prior 1993, EDGE may represent the only option to reliably estimate historical transaction costs for the U.S. stock market.

[Insert Figure 9 about here.]

Following the proliferation of electronic trading between 2001–2005, we find that

the spreads for mid and large caps have become too small to be reliably estimated from a monthly sample of daily data, as already observed by Jahan-Parvar and Zikes (2019) and Tremacoldi-Rossi and Irwin (2019). To improve the estimation accuracy for larger stocks in more recent periods, a researcher may consider using intraday price data whenever possible, as illustrated in Section 3.5.

5 Conclusion

We propose an *Efficient Discrete Generalized Estimator* (EDGE) of the bid-ask spread derived from open, high, low, and close prices. Our approach adds to the literature in two ways. First, EDGE is unbiased when trading is infrequent. Second, EDGE minimizes the estimation variance.

These properties are essential for reliable identification in applied research. We show that earlier proposed methods based on transaction prices are likely to underestimate historical spreads substantially. This is particularly evident for small stocks, where liquidity tends to be low, and transaction costs are expected to be high. In addition, the improved accuracy of EDGE reduces the probability of finding negative estimates. As negative estimates are commonly re-set to zero in empirical work, this property reduces another source of bias for all types of stocks.

We illustrate the performance of our efficient estimator in a comprehensive simulation experiment and with empirical data using the CRSP-TAQ merged database in the period 1993–2020. Our results show that EDGE generally delivers the most accurate estimates of effective bid-ask spreads numerically and empirically.

Our estimator is derived under permissive assumptions that allow for stylized facts typically observed in real market data. As such, it can be applied to a wide range of asset classes, in liquid and illiquid market segments, at low or high frequency. In particular, we show that EDGE is the first approach, relying on transaction prices only, that can be expected to work accurately when applied to high-frequency data and when the number of trading observations is sparse.

References

Abdi, F., and A. Ranaldo. 2017. A simple estimation of bid-ask spreads from daily close, high, and low prices. *Review of Financial Studies* 30:4437–80.

Amihud, Y., and H. Mendelson. 1986. Asset pricing and the bid-ask spread. *Journal of Financial Economics* 17:223–49.

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

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

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

Blankespoor, E., E. deHaan, and I. Marinovic. 2020. Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics* 70:101–344.

Bryant, H. L., and M. S. Haigh. 2004. Bid–ask spreads in commodity futures markets. *Applied Financial Economics* 14:923–36.

Corwin, S. A., and P. Schultz. 2012. A simple way to estimate bid-ask spreads from daily high and low prices. *Journal of Finance* 67:719–60.

Dechow, P. M., R. G. Sloan, and A. P. Sweeney. 1996. Causes and consequences of earnings manipulation: An analysis of firms subject to enforcement actions by the sec. *Contemporary Accounting Research* 13:1–36.

Egginton, J. F., B. F. Van Ness, and R. A. Van Ness. 2016. Quote stuffing. *Financial Management* 45:583–608.

Fang, V. W., X. Tian, and S. Tice. 2014. Does stock liquidity enhance or impede firm innovation? *Journal of Finance* 69:2085–125.

Fong, K. Y., C. W. Holden, and C. A. Trzcinka. 2017. What are the best liquidity proxies for global research? *Review of Finance* 21:1355–401.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka. 2009. Do liquidity measures measure liquidity? *Journal of Financial Economics* 92:153–81.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–54.

Hasbrouck, J. 2009. Trading costs and returns for us equities: Estimating effective costs from daily data. *Journal of Finance* 64:1445–77.

Holden, C. W., and S. Jacobsen. 2014. Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. *Journal of Finance* 69:1747–85.

Huang, R. D., and H. R. Stoll. 1994. Market integration and price execution for NYSE-listed securities. *Review of Financial Studies* 7:179–213.

Jahan-Parvar, M. R., and F. Zikes. 2019. When do low-frequency measures really measure transaction costs? Finance and Economics Discussion Series, Board of Governors of the Federal Reserve System.

Johann, T., and E. Theissen. 2017. The best in town: A comparative analysis of low-frequency liquidity estimators. *Available at SSRN 2905032* .

Karnaukh, N., A. Ranaldo, and P. Söderlind. 2015. Understanding FX liquidity. *Review of Financial Studies* 28:3073–108.

Korajczyka, R. A., and R. Sadka. 2008. Pricing the commonality across alternative measures of liquidity. *Journal of Financial Economics* 87:45–72.

Lee, C. M., B. Mucklow, and M. J. Ready. 1993. Spreads, depths, and the impact of earnings information: An intraday analysis. *The Review of Financial Studies* 6:345–74.

Lee, C. M. C. 1993. Market integration and price execution for NYSE-listed securities. *Journal of Finance* 48:1009–38.

Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka. 1999. A new estimate of transaction costs. *Review of Financial Studies* 12:1113–41.

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.

Petersen, M. A., and D. Fialkowski. 1994. Posted versus effective spread. *Journal of Financial Economics* 35:269–92.

R Core Team. 2020. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

Schestag, R., P. Schuster, and M. Uhrig-Homburg. 2016. Measuring liquidity in bond markets. *Review of Financial Studies* 29:1170–219.

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

Table 1

Generalized Bid-Ask Spread Estimation Formulas

Prices	Equations		Prices
O	$S^2 = -\frac{4Cov[o_t - o_{t-1}, o_{t-1} - o_{t-2}]}{(1-\nu_{o=o})^2}$	$S^2 = -\frac{4Cov[c_t - c_{t-1}, c_{t-1} - c_{t-2}]}{(1-\nu_{c=c})^2}$	C
OC	$S^2 = -\frac{4Cov[c_t - o_t, o_t - c_{t-1}]}{(1-\nu_{o=c})}$	$S^2 = -\frac{4Cov[o_t - c_{t-1}, c_{t-1} - o_{t-1}]}{(1-\nu_{o=c=c})(1-\nu_{o=c})}$	CO
OHL	$S^2 = -\frac{4Cov[\eta_t - o_t, o_t - \eta_{t-1}]}{(1-k\nu_{o=h,l})}$	$S^2 = -\frac{4Cov[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{(1-\nu_{h=l=c})(1-k\nu_{c=h,l})}$	CHL
OHLC	$S^2 = -\frac{4Cov[\eta_t - o_t, o_t - c_{t-1}]}{(1-k\nu_{o=h,l})}$	$S^2 = -\frac{4Cov[o_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{(1-\nu_{h=l=c})(1-k\nu_{c=h,l})}$	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 ($c_t = h_t$) or the Low ($c_t = l_t$) price respectively.

Parameters	
k	Computed as $k = 4p(1-p)$ where p is the probability of the High price to be buyer initiated or, equivalently, the probability of the Low price to be seller initiated.

Table 2

Estimated Monthly Spreads in Low Frequency

Monthly spread estimates from EDGE as proposed in this paper and the ones obtained with the estimators in Abdi and Ranaldo (2017) (AR and AR2), Corwin and Schultz (2012) (CS and CS2), and Roll (1984) for a simulated price process as described in Section 2.1.1. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of spread estimates, and the proportion of spread estimates that are nonpositive across the simulations. Panel B reports results from simulations incorporating 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	AR	AR2	CS	CS2	Roll
Panel A: Simulated Spread Estimates under Near-Ideal Conditions							
Spread 0.50%	Mean	0.44%	0.71%	1.21%	0.60%	1.44%	1.45%
	σ	0.34%	0.78%	0.36%	0.50%	0.34%	1.43%
	$\%\leq 0$	27.61%	46.44%	0.00%	19.25%	0.00%	39.50%
Spread 1.00%	Mean	0.89%	0.95%	1.32%	1.03%	1.75%	1.60%
	σ	0.44%	0.86%	0.38%	0.59%	0.38%	1.50%
	$\%\leq 0$	10.36%	35.35%	0.00%	5.62%	0.00%	36.61%
Spread 3.00%	Mean	2.92%	2.91%	2.41%	2.93%	3.22%	2.90%
	σ	0.42%	0.73%	0.51%	0.62%	0.50%	1.84%
	$\%\leq 0$	0.01%	0.80%	0.00%	0.00%	0.00%	17.54%
Spread 5.00%	Mean	4.96%	4.97%	4.32%	4.90%	4.98%	4.78%
	σ	0.41%	0.59%	0.61%	0.62%	0.58%	2.17%
	$\%\leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	6.75%
Spread 8.00%	Mean	7.98%	7.99%	7.58%	7.86%	7.86%	7.71%
	σ	0.39%	0.55%	0.58%	0.63%	0.63%	2.70%
	$\%\leq 0$	0.00%	0.00%	0.00%	0.00%	0.00%	2.31%
Panel B: Overnight Return and Only 1% Prices Observed (≈ 4 Trades per Day)							
Spread 0.50%	Mean	0.75%	0.71%	1.10%	0.02%	0.35%	1.61%
	σ	0.83%	0.80%	0.36%	0.07%	0.15%	1.59%
	$\%\leq 0$	46.53%	47.84%	0.00%	86.53%	0.00%	39.27%
Spread 1.00%	Mean	0.99%	0.86%	1.19%	0.03%	0.40%	1.74%
	σ	0.91%	0.86%	0.38%	0.09%	0.17%	1.64%
	$\%\leq 0$	36.76%	41.50%	0.00%	82.33%	0.00%	36.95%
Spread 3.00%	Mean	2.87%	2.22%	1.99%	0.28%	0.85%	2.91%
	σ	0.96%	1.06%	0.54%	0.33%	0.30%	1.98%
	$\%\leq 0$	2.92%	9.52%	0.00%	38.81%	0.00%	20.43%
Spread 5.00%	Mean	5.04%	4.01%	3.23%	0.99%	1.56%	4.67%
	σ	0.84%	0.98%	0.73%	0.62%	0.50%	2.33%
	$\%\leq 0$	0.05%	0.64%	0.00%	6.94%	0.00%	9.10%
Spread 8.00%	Mean	8.02%	6.58%	5.33%	2.46%	2.86%	7.50%
	σ	0.88%	1.03%	1.03%	0.97%	0.86%	2.87%
	$\%\leq 0$	0.00%	0.01%	0.00%	0.33%	0.00%	3.37%

Table 3

Summary Statistics

The table reports the number of spread estimates, the mean, median, and standard deviation for the estimates and for the effective spread benchmark. The correlation with the effective spread benchmark, the root mean squared error (RMSE), the mean absolute percentage error (MAPE) are also reported, together with the proportion of spread estimates that are nonpositive. The sample period is from 1993–2020 (CRSP-TAQ merged sample).

Estimator:	N	Mean	Median	Sd	Cor	MAPE	RMSE	$\% \leq 0$
Units:	1	%	%	%	%	%	1	%
EDGE	1,626,448	2.23	1.05	3.58	**76.48**	**16.95**	**1.22**	**24.79**
AR	1,626,448	1.75	0.96	2.58	66.87	20.07	1.39	31.65
AR2	1,626,448	1.70	1.18	1.71	64.57	22.12	1.46	–
CS	1,626,448	0.68	0.28	1.17	45.60	33.95	2.08	29.22
CS2	1,626,448	1.32	0.94	1.34	43.60	33.44	2.35	–
Roll	1,626,448	2.67	1.36	48.09	4.91	24.85	1.77	32.54
ES Benchmark	1,626,448	1.88	0.76	2.99	–	–	–	–

Table 4

Correlation with Monthly TAQ Effective Spreads

The table shows group specific correlations of spread estimates with the TAQ effective spread. The table also reports the median effective spread per group and the fraction of spread estimates that are non-positive. 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, AR and AR2 are the estimators proposed by Abdi and Ranaldo (2017), CS and CS2 are the estimators proposed by Corwin and Schultz (2012), and the Roll (1984) estimator. All estimators are based on daily observations using a monthly estimation window. The sample period is from 1993–2020 (CRSP-TAQ merged sample).

		Correlation (%)						$\%\leq 0$			
Group	Spread	**EDGE**	AR	AR2	CS	CS2	Roll	**EDGE**	AR	CS	Roll
Panel A: Analysis across different markets											
NYSE	0.16%	**57**	43	47	42	42	1	**40**	44	42	**40**
AMEX	1.75%	**68**	62	62	42	44	11	**26**	32	41	33
NASDAQ	1.38%	**76**	66	62	42	38	8	**17**	26	22	29
Panel B: Analysis across time periods											
1993–1996	2.49%	**83**	76	69	48	48	49	**15**	23	26	26
1997–2000	1.68%	**78**	69	67	48	47	34	**22**	30	35	32
2001–2002	1.25%	**74**	69	68	45	45	14	**24**	31	35	32
2003–2007	0.31%	**64**	55	59	34	36	8	**26**	34	29	35
2008–2011	0.25%	**62**	52	51	30	29	1	**26**	32	27	33
2012–2015	0.18%	**55**	47	49	30	28	5	31	37	**25**	36
2016–2020	0.18%	**53**	40	44	36	33	5	34	39	**28**	35
Panel C: Analysis across market capitalization											
Quintile 1	3.14%	**71**	63	60	39	37	21	**15**	23	25	26
Quintile 2	2.09%	**69**	57	51	32	25	11	**16**	23	26	26
Quintile 3	1.08%	**72**	56	51	35	27	3	**20**	27	26	31
Quintile 4	0.30%	**75**	56	55	45	40	12	**31**	37	**31**	37
Quintile 5	0.09%	**51**	36	40	37	36	0	**38**	44	**38**	40
Panel D: Analysis across spread sizes											
Quintile 1	0.08%	18	14	**23**	12	21	0	41	46	**38**	41
Quintile 2	0.26%	**45**	31	38	32	34	5	**34**	41	**34**	39
Quintile 3	0.75%	**62**	46	50	42	40	3	**24**	33	28	36
Quintile 4	1.81%	**67**	55	55	39	36	8	**15**	24	24	29
Quintile 5	4.38%	**66**	59	55	34	34	22	**10**	16	22	19

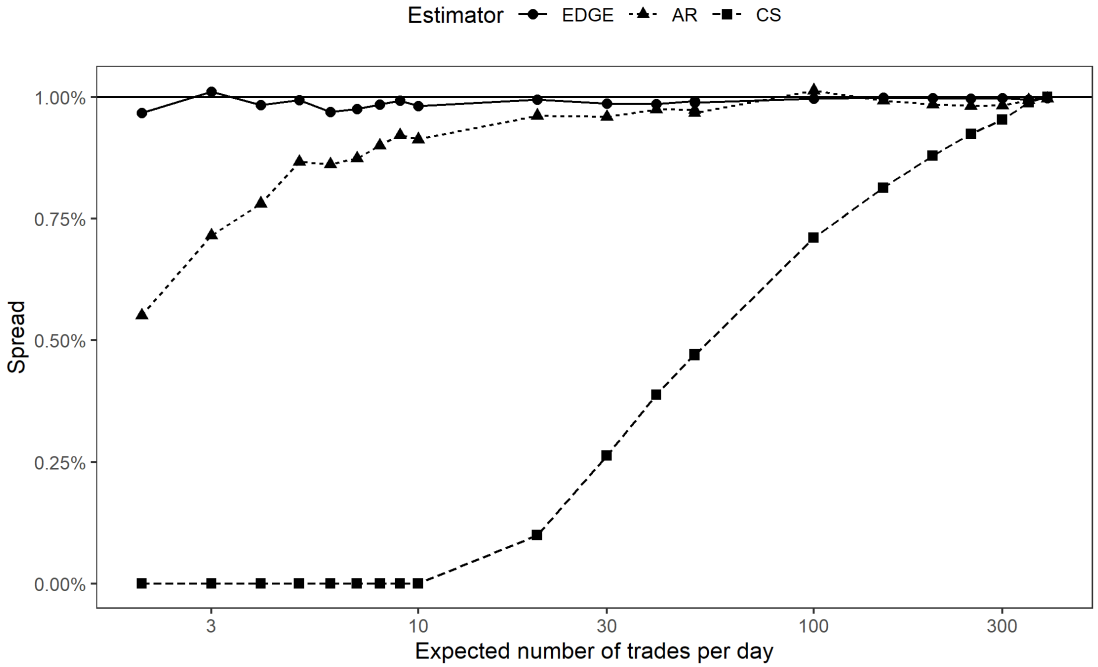


Figure 1: Comparison of bid-ask spread estimates based on EDGE as proposed in this paper with the estimators by Corwin and Schultz (2012) (CS) and Abdi and Ranaldo (2017) (AR), for a simulated price process as described in Section 2.2.1. The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per day is specified in the horizontal axis. The simulations use a constant spread of 1%.

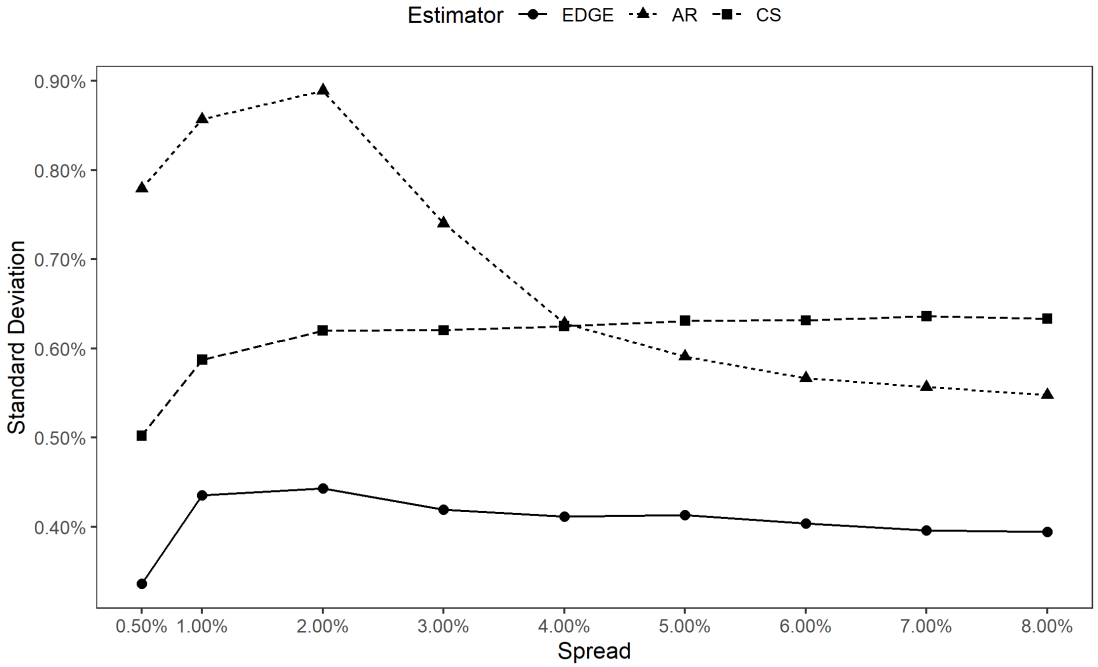


Figure 2: Comparison of the standard deviation of bid-ask spread estimates based on EDGE as proposed in this paper with the estimators by Corwin and Schultz (2012) (CS) and Abdi and Ranaldo (2017) (AR), for several spread levels (horizontal axis) as described in Section 2.2.2. These simulations use 390 trades per day, so that all the estimators are unbiased (see Figure 1) and the minimum-variance estimator coincides with the best estimator in the usual root mean squared error sense.

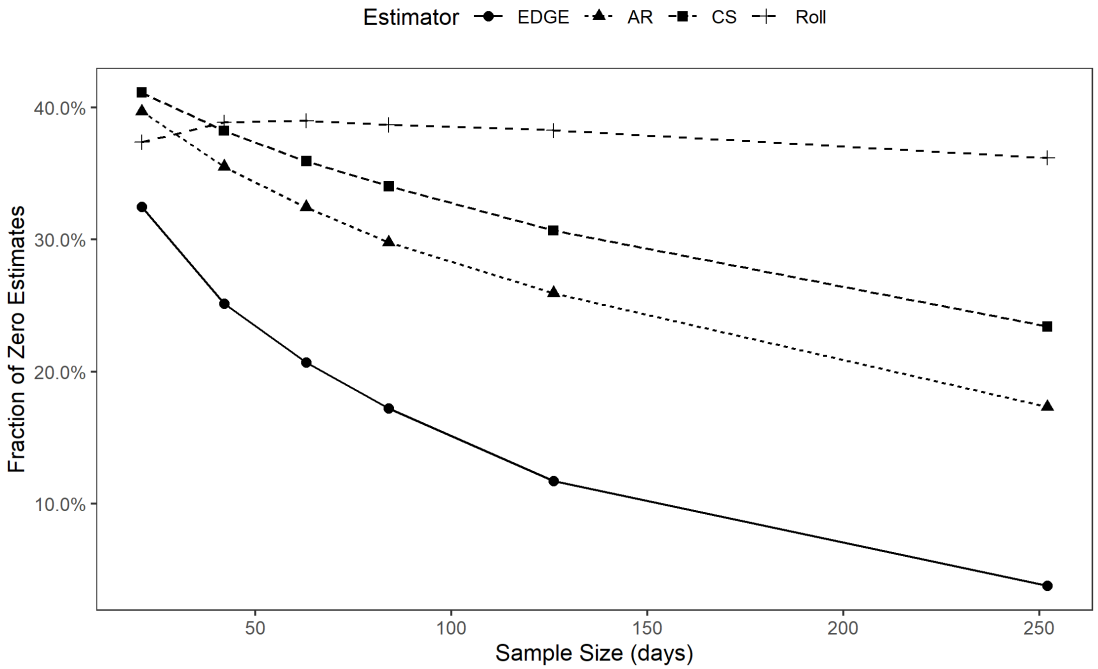


Figure 3: Proportion of zero bid-ask spreads estimates based on EDGE as proposed in this paper, as well as the estimators proposed by Corwin and Schultz (2012) (CS), Abdi and Ranaldo (2017) (AR), and Roll (1984) for sample sizes ranging from one month to one year (horizontal axis) as described in Section 2.2.3. 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.

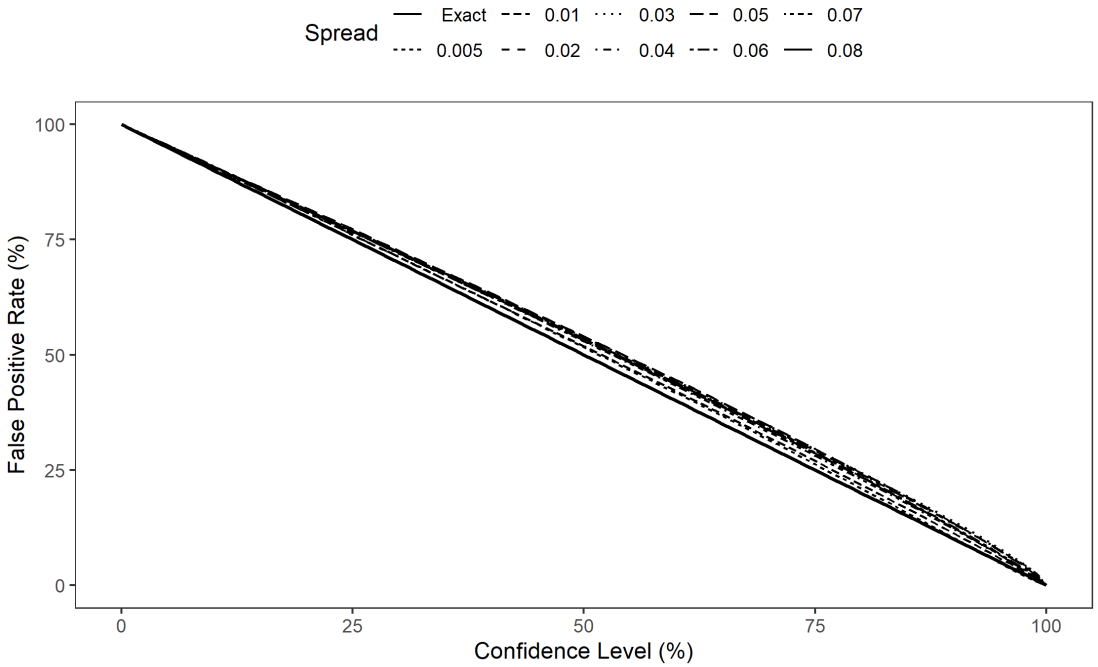
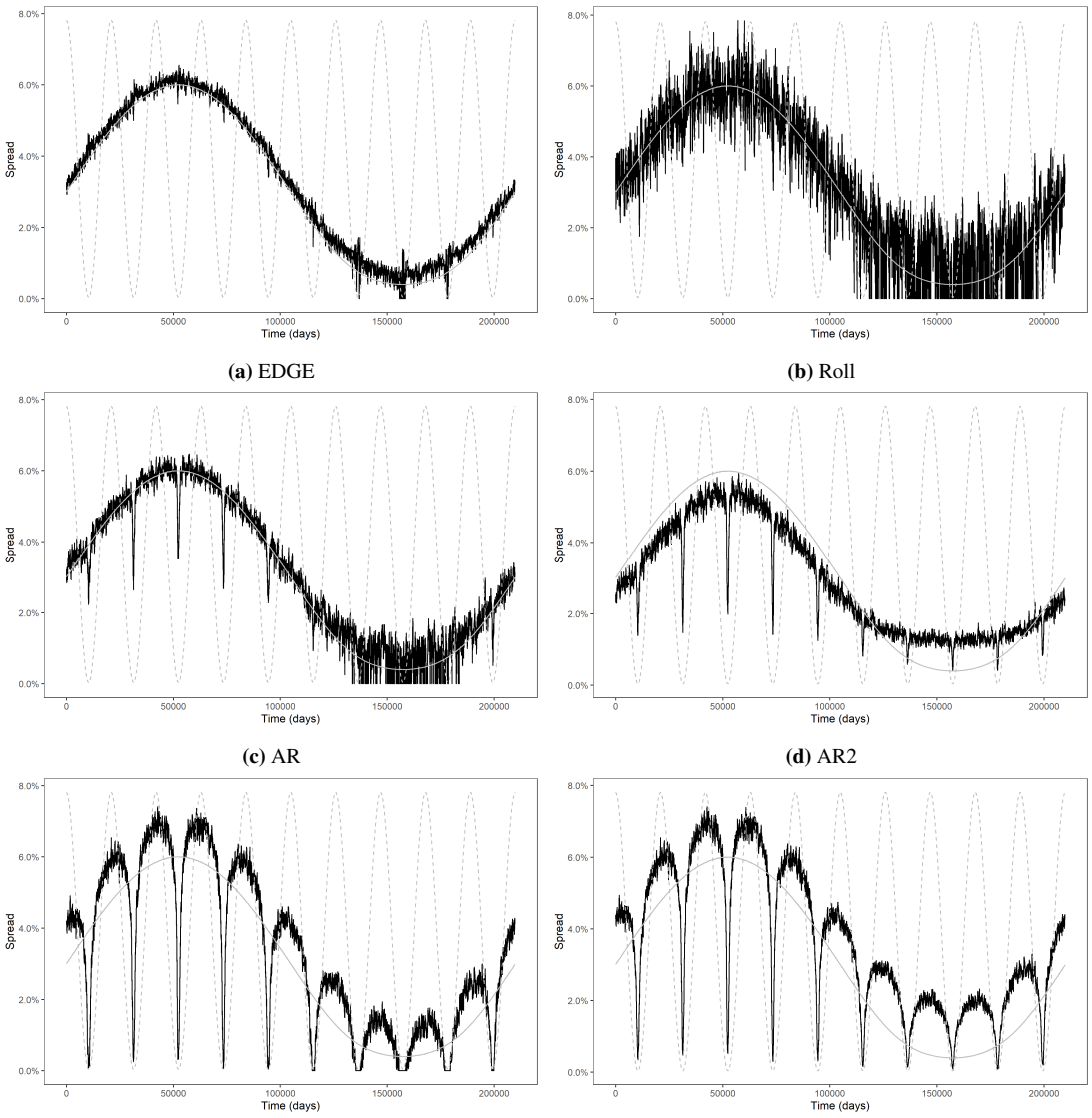


Figure 4: False positive rates (vertical axis) against confidence levels (horizontal axis) from our model in Equation (20) for several spread levels as described in Section 2.2.4. As with a 95% confidence level we expect 5% false positives, the exact theoretical relationship is $y = 1 - x$ (solid line in black). These simulations use 390 trades per day.



(a) EDGE **(b)** Roll

(c) AR **(d)** AR2

(e) CS **(f)** CS2

Figure 5: This figure shows the time-series estimates from EDGE as proposed in this paper and the ones obtained with the estimators in Abdi and Ranaldo (2017) (AR and AR2), Corwin and Schultz (2012) (CS and CS2), and Roll (1984) 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 $p = 0.5 + 0.495 \times \cos(\frac{20\pi t}{n})$ 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 $\mu = 0.03 \times (1 + \sin(\frac{2\pi t}{n}))$. Then, for each minute the spread is randomly drawn from a normal distribution with mean μ 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 spread (black line). The estimates are benchmarked with the average spread (solid line in grey) and the average (scaled) probability of observing a trade (dotted line in grey) in the previous year.

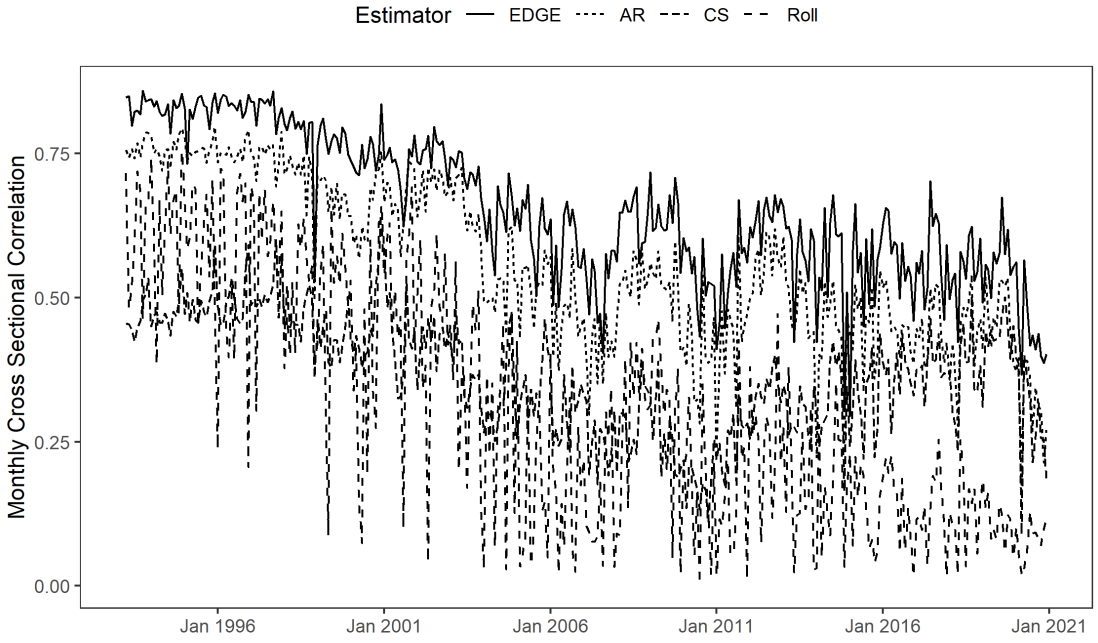


Figure 6: Month-by-month cross-sectional correlations with the TAQ benchmark in Equation (22) for various spread estimators as described in Section 3.2.

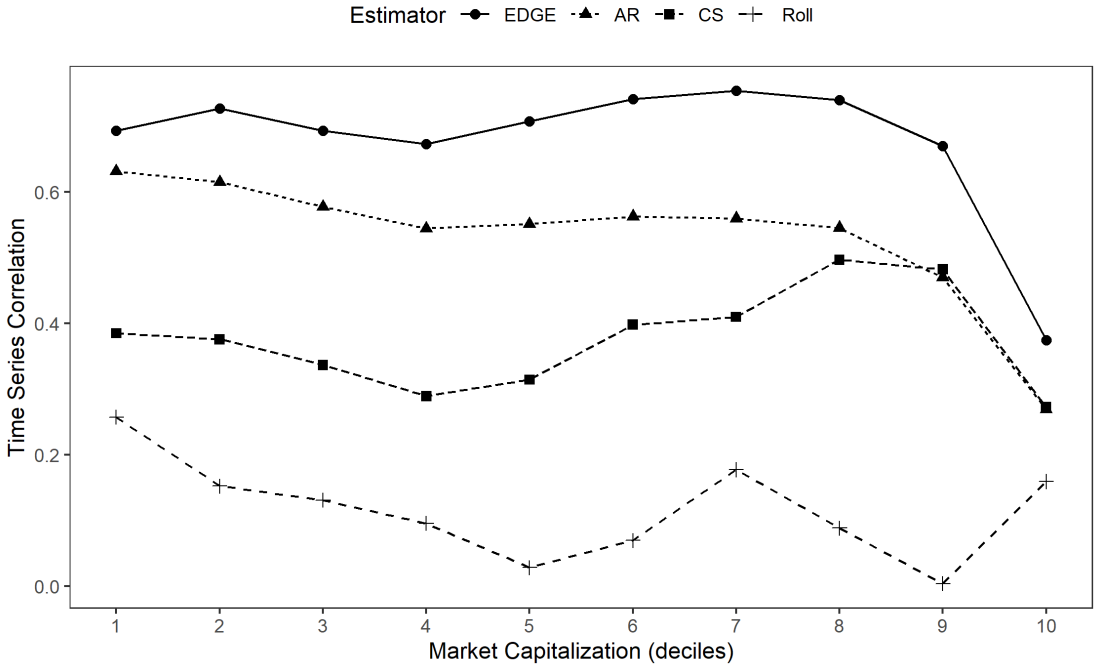
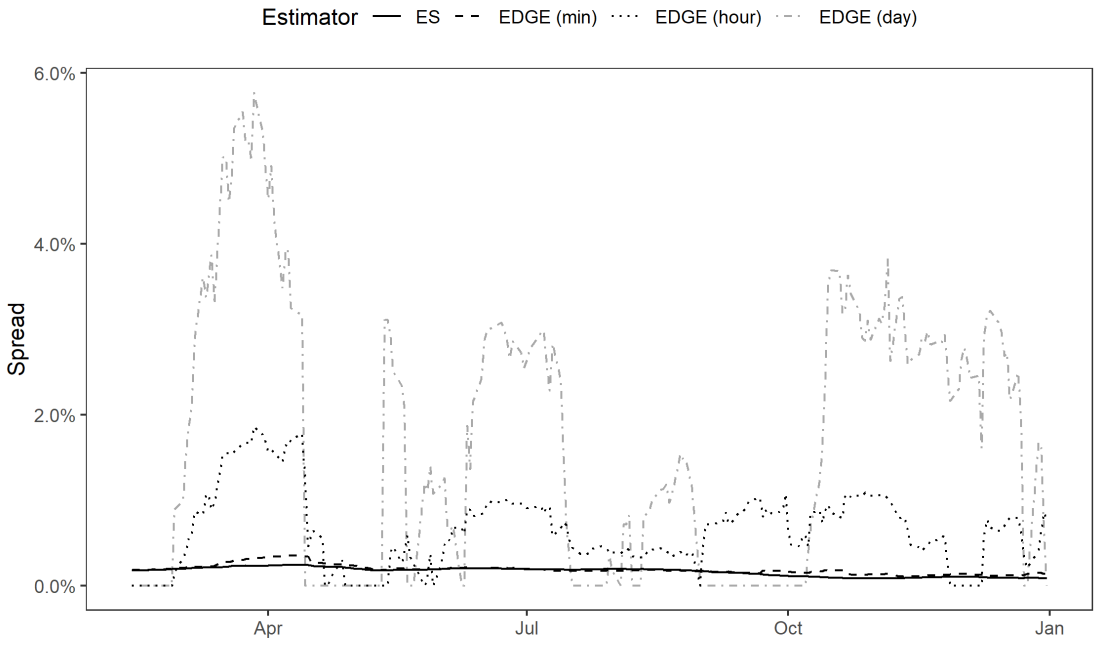
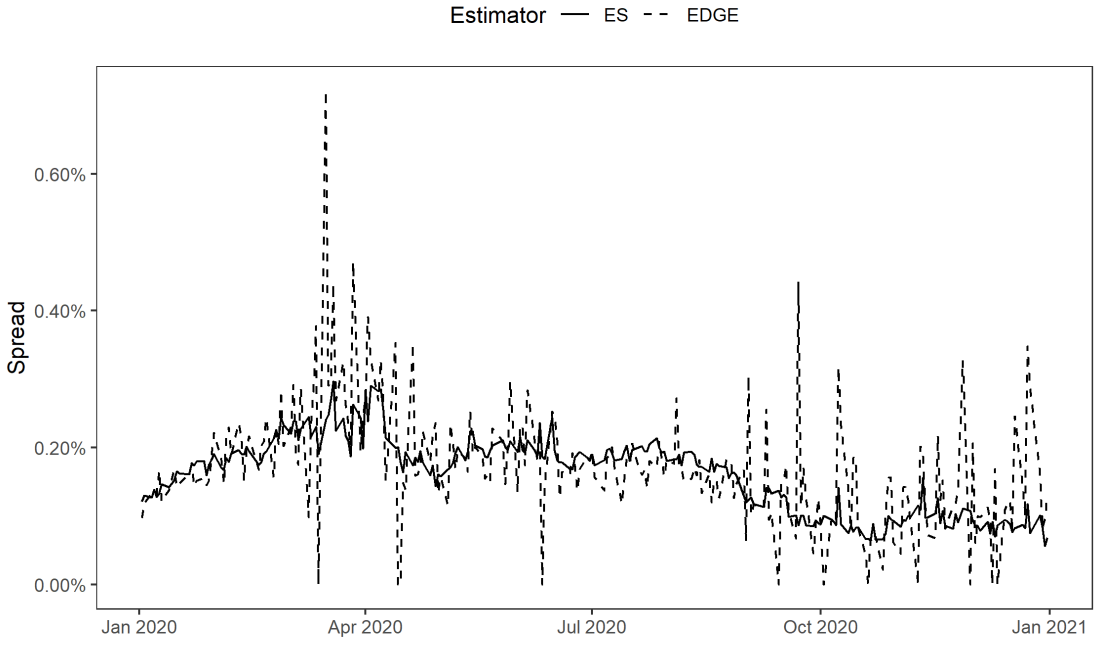


Figure 7: Time-series correlations for deciles sorted on size with the TAQ benchmark in Equation (22) obtained using various spread estimators as described in Section 3.3.

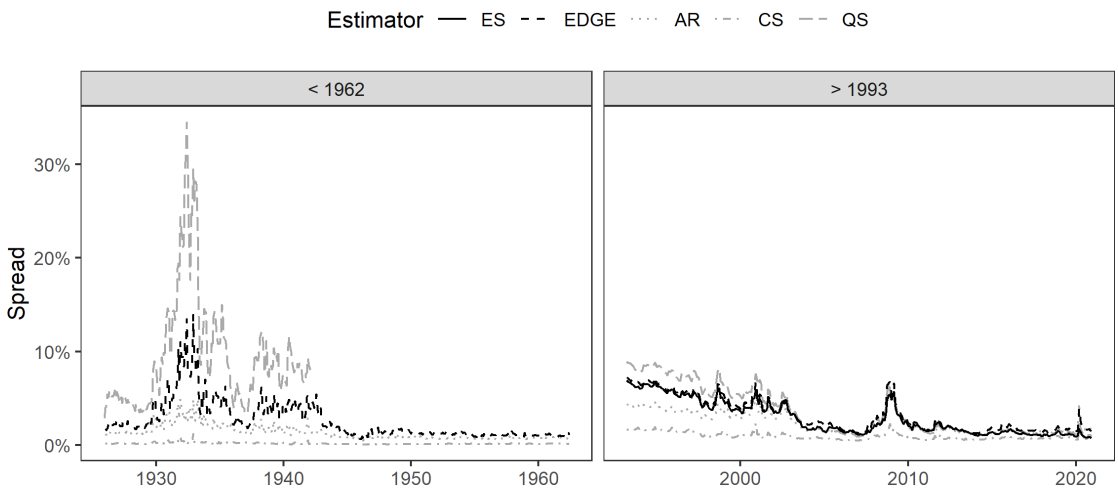


(a) Monthly estimates

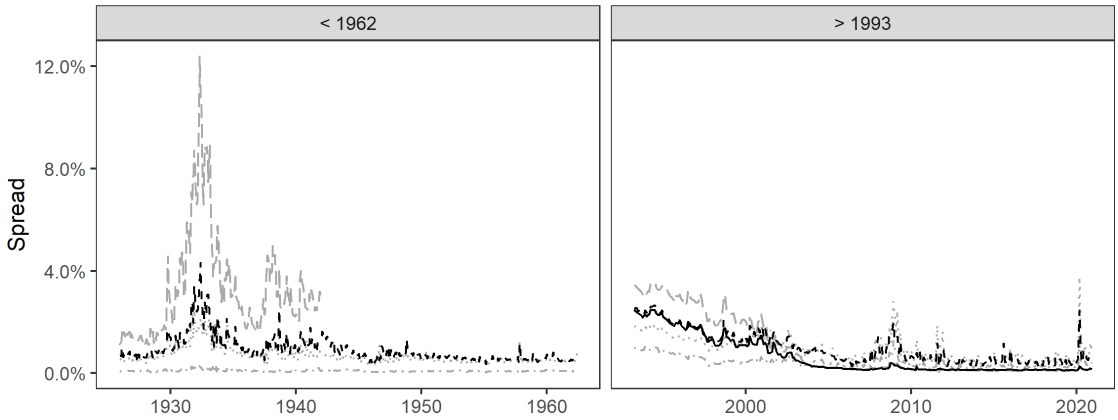


(b) Daily estimates

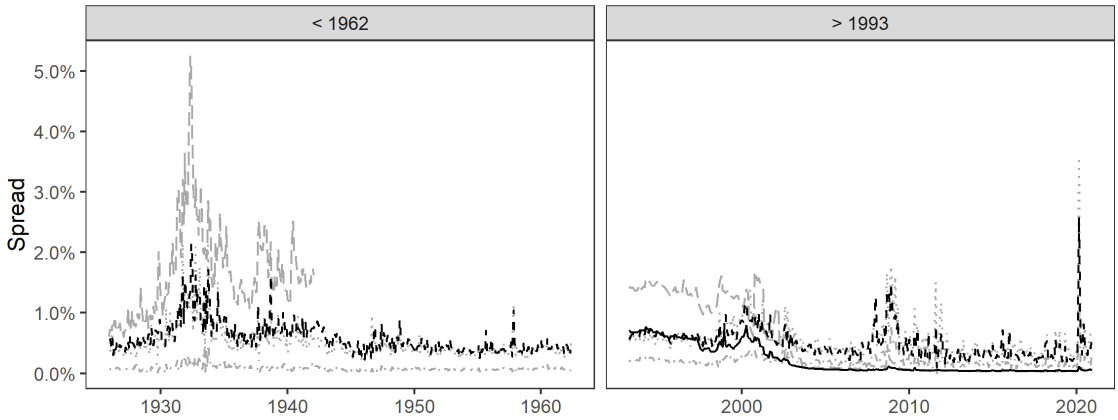
Figure 8: Spread estimates for GameStop Corp. (GME) in 2020. Figure (a) reports the monthly estimates (21-day rolling window) obtained using daily, hourly, or minute price data, together with the average effective spread benchmark in the corresponding time window. Figure (b) reports the daily estimates obtained from intraday minute data and the corresponding effective spread benchmark within the day.



(a) Small Cap Stocks



(b) Mid Cap Stocks



(c) Large Cap Stocks

Figure 9: The graphs plot the time series of spread estimates of several methods as well as the effective spread benchmark for (a) small caps (b) mid caps and (c) large caps as described in Section 4. Historical quoted spreads from CRSP are also reported for the periods 1925–1942 and 1993–2020. CRSP quote data are not available between 1942–1993.

A Appendix

A.1 Moments of Z_t

To compute the moments of $Z_t = S(B_t - 0.5)$ we compute its moment generating function (MGF). The MGF of the Bernoulli random variable B with probability of success p is:

$$M_B(t) = (1-p) + pe^t \,.$$

Since Z_t is a linear transformation of B_t, its MGF can be obtained from the MGF of B_t:

$$M_Z(t) = (1-p)e^{-\frac{S}{2}t} + pe^{\frac{S}{2}t} \,.$$

The moments are computed by differentiation:

$$\mathbb{E}[Z_t^n] = \frac{d^n M_Z(t)}{dt^n}\Big|_{t=0} = (1-p)\Big(-\frac{S}{2}\Big)^n + p\Big(\frac{S}{2}\Big)^n = \begin{cases} \Big(\frac{S}{2}\Big)^n (2p-1) & n = 1,3,5,... \\ \Big(\frac{S}{2}\Big)^n & n = 2,4,6,... \end{cases}$$

And in particular, we have:

$$\mathbb{E}[Z_t] = \frac{S}{2}(2p-1)\,, \quad \mathbb{E}[Z_t^2] = \frac{S^2}{4}\,, \quad \mathbb{V}[Z_t] = S^2 p(1-p)\,.$$

That for $p = 0.5$ become:

$$\mathbb{E}[Z_t] = 0\,, \quad \mathbb{E}[Z_t^2] = \frac{S^2}{4}\,, \quad \mathbb{V}[Z_t] = \frac{S^2}{4}\,.$$

A.1.1 Random Spread

When considering a random spread S_t, we compute:

$$\begin{aligned} \mathbb{V}[Z_t] &= \mathbb{V}[S_t(B_t - 0.5)] \\ &= \mathbb{E}[S_t^2(B_t - 0.5)^2] - \mathbb{E}[S_t(B_t - 0.5)]^2 \\ &= \mathbb{E}[S_t^2]\mathbb{E}[B_t^2 - B_t + 0.25] - \mathbb{E}[S_t]\mathbb{E}[B_t - 0.5] \\ &= \frac{\mathbb{E}[S_t^2]}{4}\,. \end{aligned}$$

A.2 The Generalized Estimators

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

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

In the particular case when $\mathbb{E}[X \mid Z] = 0$ or $\mathbb{E}[Y \mid Z] = 0$, we have:

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}[\mathbb{C}\text{ov}[X, Y \mid Z]] \,. \tag{A.1}$$

A.2.1 C prices

We need to compute the covariance:

$$\mathbb{C}\text{ov}[c_t - c_{t-1}, c_{t-1} - c_{t-2}] \,.$$

We replace the observed log-prices c_t with the actual (but unobserved) log-prices $\tilde{c}_t$ by Equation (2) and expand the covariance in the four terms:

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

where the first three terms are zero since the actual returns are uncorrelated and independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned} \mathbb{C}\text{ov}[Z_t - Z_{t-1}, Z_{t-1} - Z_{t-2}] &= \mathbb{C}\text{ov}[Z_t, Z_{t-1}] \\ &+ \mathbb{C}\text{ov}[Z_t, -Z_{t-2}] \\ &+ \mathbb{C}\text{ov}[-Z_{t-1}, Z_{t-1}] \\ &+ \mathbb{C}\text{ov}[-Z_{t-1}, -Z_{t-2}] \,. \end{aligned} \tag{A.3}$$

Since the random variables Z are independent for different trades, we might assume that the only non-vanishing term is $\mathbb{C}\text{ov}[-Z_{t-1}, Z_{t-1}] = -\mathbb{V}[Z]$. However, we should pay extra care when no trade is observed for period t. In this case the market reports the previous closing price so that Z_t and Z_{t-1} are generated by the same trade. In this case $\mathbb{C}\text{ov}[Z_t, Z_{t-1}] = \mathbb{V}[Z]$. By decomposing the covariance with Equation (A.1), we have:

$$\begin{aligned} \mathbb{C}\text{ov}[Z_t, Z_{t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_t, Z_{t-1} | Z_t = Z_{t-1}]] \\ &= \mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}] \,, \end{aligned}$$

where $\mathbb{P}[Z_t = Z_{t-1}]$ is the probability that the same trade generated both Z_t and Z_{t-1}. The same holds for:

$$\begin{aligned} \mathbb{C}\text{ov}[-Z_{t-1}, -Z_{t-2}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{t-1}, Z_{t-2} | Z_{t-1} = Z_{t-2}]] \\ &= \mathbb{V}[Z]\mathbb{P}[Z_{t-1} = Z_{t-2}] \,, \end{aligned}$$

where $\mathbb{P}[Z_{t-1} = Z_{t-2}]$ is the probability that the same trade generated both Z_{t-1} and Z_{t-2}. Moreover, we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_t, -Z_{t-2}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_t, -Z_{t-2}|Z_t = Z_{t-2}]] \\ &= -\mathbb{V}[Z]\mathbb{P}[Z_t = Z_{t-1}]\mathbb{P}[Z_{t-1} = Z_{t-2}]\,.\end{aligned}$$

We estimate the probability that two subsequent prices are generated by the same trade by counting the fraction of times, $\nu_{c=c}$, for which the closing prices over two subsequent time periods are equal.

$$\mathbb{P}[Z_t = Z_{t-1}] = \mathbb{P}[Z_{t-1} = Z_{t-2}] \hat{=} \nu_{c=c}\,.$$

By considering Equation (A.2) and rewriting Equation (A.3), we have:

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

The final formula is obtained by computing the variance of Z in Appendix A.1.

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

A.2.2 O prices

By replacing c_t with o_t and following the same steps illustrated in Section A.2.1, we obtain:

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

A.2.3 CO prices

We need to compute the covariance:

$$\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}]\,.$$

We replace the observed log-prices with the actual (but unobserved) log-prices by Equation (2) and expand the covariance in the four terms:

$$\begin{aligned}\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}] &= \mathbb{C}\text{ov}[\tilde{o}_t - \tilde{c}_{t-1}, \tilde{c}_{t-1} - \tilde{o}_{t-1}] \\ &+ \mathbb{C}\text{ov}[\tilde{o}_t - \tilde{c}_{t-1}, Z_{c,t-1} - Z_{o,t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{o,t} - Z_{c,t-1}, \tilde{c}_{t-1} - \tilde{o}_{t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{o,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{o,t-1}] \\ &= \mathbb{C}\text{ov}[Z_{o,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{o,t-1}]\,.\end{aligned} \tag{A.4}$$

where the first three terms are zero since the actual returns are uncorrelated and independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{o,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{o,t-1}] = \mathbb{C}\text{ov}[Z_{o,t}, Z_{c,t-1}] \\ + \mathbb{C}\text{ov}[Z_{o,t}, -Z_{o,t-1}] \\ + \mathbb{C}\text{ov}[-Z_{c,t-1}, Z_{c,t-1}] \\ + \mathbb{C}\text{ov}[-Z_{c,t-1}, -Z_{o,t-1}]\,.\end{aligned} \tag{A.5}$$

Since the random variables Z are independent for different trades, we might assume that the only non-vanishing term is $\mathbb{C}\text{ov}[-Z_{c,t-1}, Z_{c,t-1}] = -\mathbb{V}[Z]$. However, we should pay extra care when no trade is observed for period t and when only a single trade is observed for period $t-1$. In this first case, the market reports the previous closing price so that $Z_{o,t}$ and $Z_{c,t-1}$ are generated by the same trade. In the second case, $Z_{c,t-1}$ and $Z_{o,t-1}$ are generated by the same trade. In both cases, the covariance reduces to $\mathbb{V}[Z]$. By decomposing the covariance with Equation (A.1), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{o,t}, Z_{c,t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{o,t}, Z_{c,t-1}|Z_{o,t} = Z_{c,t-1}]] \\ &= \mathbb{V}[Z]\mathbb{P}[Z_{o,t} = Z_{c,t-1}]\,,\end{aligned}$$

where $\mathbb{P}[Z_{o,t} = Z_{c,t-1}]$ is the probability that the opening price and the previous close are generated by the same trade. The same holds for:

$$\begin{aligned}\mathbb{C}\text{ov}[-Z_{c,t-1}, -Z_{o,t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{o,t-1}|Z_{c,t-1} = Z_{o,t-1}]] \\ &= \mathbb{V}[Z]\mathbb{P}[Z_{c,t-1} = Z_{o,t-1}]\,,\end{aligned}$$

where $\mathbb{P}[Z_{c,t-1} = Z_{o,t-1}]$ is the probability that the open and close price in the same period are generated by the same trade. Moreover, we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{o,t}, -Z_{o,t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{o,t}, -Z_{o,t-1}|Z_{o,t} = Z_{o,t-1}]] \\ &= -\mathbb{V}[Z]\mathbb{P}[Z_{o,t} = Z_{c,t-1}]\mathbb{P}[Z_{c,t-1} = Z_{o,t-1}]\,.\end{aligned} \tag{A.6}$$

We estimate the probability that the opening price and the previous close are generated by the same trade by counting the fraction of times $\nu_{o=c=c}$ in which both the closing and the opening prices are equal to the previous close. Moreover, we estimate the probability that the open and close price in the same period are generated by the same trade by counting the fraction of times $\nu_{o=c}$ in which the opening and closing prices are equal.

$$\mathbb{P}[Z_{o,t} = Z_{c,t-1}] \mathrel{\hat{=}} \nu_{o=c=c}\,, \quad \mathbb{P}[Z_{c,t-1} = Z_{o,t-1}] \mathrel{\hat{=}} \nu_{o=c}\,.$$

By considering Equation (A.4) and rewriting Equation (A.5), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - o_{t-1}] &= -\mathbb{V}[Z](1 - \nu_{o=c=c} - \nu_{o=c} + \nu_{o=c=c}\nu_{o=c}) \\ &= -\mathbb{V}[Z](1 - \nu_{o=c=c})(1 - \nu_{o=c})\,.\end{aligned}$$

The final formula is obtained by computing the variance of Z in Appendix A.1.

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

A.2.4 OC prices

We need to compute the covariance:

$$\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}]\,.$$

We replace the observed log-prices with the actual (but unobserved) log-prices by Equation (2) and expand the covariance in the four terms:

$$\begin{aligned}\mathbb{C}\text{ov}[c_t - o_t, o_t - c_{t-1}] &= \mathbb{C}\text{ov}[\tilde{c}_t - \tilde{o}_t, \tilde{o}_t - \tilde{c}_{t-1}] \\ &+ \mathbb{C}\text{ov}[\tilde{c}_t - \tilde{o}_t, Z_{o,t} - Z_{c,t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, \tilde{o}_t - \tilde{c}_{t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, Z_{o,t} - Z_{c,t-1}] \\ &= \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, Z_{o,t} - Z_{c,t-1}]\,,\end{aligned} \tag{A.7}$$

where the first three terms are zero since the actual returns are uncorrelated and independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, Z_{o,t} - Z_{c,t-1}] &= \mathbb{C}\text{ov}[Z_{c,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[-Z_{o,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, -Z_{c,t-1}]\,.\end{aligned} \tag{A.8}$$

Since the random variables Z are independent for different trades, we might assume that the only non-vanishing term is $\mathbb{C}\text{ov}[-Z_{o,t}, Z_{o,t}] = -\mathbb{V}[Z]$. However, we should pay extra care when at most one trade is observed for period t. In this case, $Z_{c,t}$ and $Z_{o,t}$ are generated by the same trade and their covariance reduces to $\mathbb{V}[Z]$. By decomposing the covariance with Equation (A.1), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{c,t}, Z_{o,t}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{c,t}, Z_{o,t}|Z_{c,t} = Z_{o,t}]] \\ &= \mathbb{V}[Z]\mathbb{P}[Z_{c,t} = Z_{o,t}]\,,\end{aligned}$$

where $\mathbb{P}[Z_{c,t} = Z_{o,t}]$ is the probability that the open and close price in the same period are generated by the same trade. The last term left to compute is $\mathbb{C}\text{ov}[Z_{c,t} - Z_{o,t}, -Z_{c,t-1}]$. This 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) is no trade is observed for period t then $Z_{c,t} - Z_{o,t} = 0$.
We estimate the probability that the open and close price in the same period are generated by the same trade by counting the fraction of times in which the close and open prices are equal.

$$\mathbb{P}[Z_{c,t} = Z_{o,t}] \hat{=} \nu_{o=c}\,.$$

By considering Equation (A.8) and rewriting Equation (A.7), we have:

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

The final formula is obtained by computing the variance of Z in Appendix A.1.

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

A.2.5 CHL prices

Let us define:

$$\eta_t = \frac{h_t + l_t}{2}, \quad Z_\eta = \frac{Z_{h,t} + Z_{l,t}}{2}.$$

We need to compute the covariance:

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

We replace the observed log-prices with the actual (but unobserved) log-prices by Equation (2) and expand the covariance in the four terms:

$$\begin{aligned}
\mathbb{C}\text{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}] &= \mathbb{C}\text{ov}[\tilde{\eta}_t - \tilde{c}_{t-1}, \tilde{c}_{t-1} - \tilde{\eta}_{t-1}] \\
&\quad + \mathbb{C}\text{ov}[\tilde{\eta}_t - \tilde{c}_{t-1}, Z_{c,t-1} - Z_{\eta,t-1}] \\
&\quad + \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{c,t-1}, \tilde{c}_{t-1} - \tilde{\eta}_{t-1}] \\
&\quad + \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{\eta,t-1}] \\
&= \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{\eta,t-1}],
\end{aligned} \tag{A.9}$$

where the first three terms are zero since the actual returns are uncorrelated and independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned}
\mathbb{C}\text{ov}[Z_{\eta,t} - Z_{c,t-1}, Z_{c,t-1} - Z_{\eta,t-1}] &= \mathbb{C}\text{ov}[Z_{\eta,t}, Z_{c,t-1}] \\
&\quad + \mathbb{C}\text{ov}[Z_{\eta,t}, -Z_{\eta,t-1}] \\
&\quad + \mathbb{C}\text{ov}[-Z_{c,t-1}, Z_{c,t-1}] \\
&\quad + \mathbb{C}\text{ov}[-Z_{c,t-1}, -Z_{\eta,t-1}].
\end{aligned} \tag{A.10}$$

Since the random variables Z are independent for different trades, we might assume that the only non-vanishing term is $\mathbb{C}\text{ov}[-Z_{c,t-1}, Z_{c,t-1}] = -\mathbb{V}[Z]$. However, we should pay extra care when no trade is observed for period t and when the closing price is selected as the high or low price for period $t-1$. In this first case, the market reports the previous closing price so that $Z_{\eta,t}$ and $Z_{c,t-1}$ are generated by the same trade and their covariance reduces to $\mathbb{V}[Z]$. In the second case $Z_{c,t-1} = Z_{h,t-1}$ and/or $Z_{c,t-1} = Z_{l,t-1}$, so that $\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{\eta,t-1}] \neq 0$. By decomposing the covariance with Equation (A.1), we have:

$$\begin{aligned}
\mathbb{C}\text{ov}[Z_{\eta,t}, Z_{c,t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{\eta,t}, Z_{c,t-1} | Z_{\eta,t} = Z_{c,t-1}]] \\
&= \mathbb{V}[Z]\mathbb{P}[Z_{\eta,t} = Z_{c,t-1}],
\end{aligned}$$

where $\mathbb{P}[Z_{\eta,t} = Z_{c,t-1}]$ is the probability that the high, low, and previous close prices are generated by the same trade. Moreover, we have:

$$\begin{aligned}
\mathbb{C}\text{ov}[-Z_{c,t-1}, -Z_{\eta,t-1}] &= 0.5(\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{h,t-1} + Z_{l,t-1}]) \\
&= 0.5(\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{h,t-1}] + \mathbb{C}\text{ov}[Z_{c,t-1}, Z_{l,t-1}]) \\
&= 0.5\mathbb{E}[\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{h,t-1} \mid Z_{c,t-1} = Z_{h,t-1}]] \\
&\quad + 0.5\mathbb{E}[\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{l,t-1} \mid Z_{c,t-1} = Z_{l,t-1}]] \\
&= 0.5\mathbb{V}[Z_p]\big(\mathbb{P}[Z_{c,t-1} = Z_{h,t-1}] + \mathbb{P}[Z_{c,t-1} = Z_{l,t-1}]\big),
\end{aligned} \tag{A.11}$$

where $\mathbb{P}[Z_{c,t-1} = Z_{h,t-1}]$ and $\mathbb{P}[Z_{c,t-1} = Z_{l,t-1}]$ are the probabilities that the closing price is selected as the high or low price respectively, and where the variance of Z_p depends on the probability p of the high price to be buyer initiated or, equivalently, of the low price to be seller initiated. The last term left to compute is:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{\eta,t}, -Z_{\eta,t-1}] &= \mathbb{E}[\mathbb{C}\text{ov}[Z_{\eta,t}, -Z_{\eta,t-1}|Z_{\eta,t} = Z_{c,t-1}]] \\ &= -\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{\eta,t-1}]\mathbb{P}[Z_{\eta,t} = Z_{c,t-1}]\,,\end{aligned} \tag{A.12}$$

where $\mathbb{C}\text{ov}[Z_{c,t-1}, Z_{\eta,t-1}]$ is given in Equation (A.11)
We estimate the probability that the high, low, and previous close prices are generated by the same trade 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 $t-1$.

$$\mathbb{P}[Z_{\eta,t} = Z_{c,t-1}] \hat{=} \nu_{h=l=c}\,.$$

Moreover, we estimate the probability that the closing price is selected as the high or low price by counting the fraction of times in which the closing price matches the high ($\nu_{c=h}$) or low ($\nu_{c=l}$) price:

$$\mathbb{P}[Z_{c,t-1} = Z_{h,t-1}] \hat{=} \nu_{c=h}\,, \quad \mathbb{P}[Z_{c,t-1} = Z_{l,t-1}] \hat{=} \nu_{c=l}\,.$$

By considering Equation (A.9) and rewriting Equation (A.10), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}] &= -\mathbb{V}[Z] \\ &\quad + \mathbb{V}[Z]\nu_{h=l=c} \\ &\quad + 0.5\mathbb{V}[Z_p](\nu_{c=h} + \nu_{c=l}) \\ &\quad - 0.5\mathbb{V}[Z_p](\nu_{c=h} + \nu_{c=l})\nu_{h=l=c} \\ &= -\mathbb{V}[Z](1 - \nu_{h=l=c})(1 - k(\nu_{c=h} + \nu_{c=l})/2)\,.\end{aligned}$$

where $k = \mathbb{V}[Z_p]/\mathbb{V}[Z] = 4p(1-p)$ is the ratio between the variance of Z_p with a generic probability p and the variance of Z with $p = \frac{1}{2}$. The final formula is obtained by computing the variance of Z in Appendix A.1:

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

A.2.6 OHL prices

Let us define:

$$\eta_t = \frac{h_t + l_t}{2}, \quad Z_\eta = \frac{Z_{h,t} + Z_{l,t}}{2}\,.$$

We need to compute the covariance:

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

We replace the observed log-prices with the actual (but unobserved) log-prices by Equation (2) and expand the covariance in the four terms:

$$\begin{aligned}\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}] &= \mathbb{C}\text{ov}[\tilde{\eta}_t - \tilde{o}_t, \tilde{o}_t - \tilde{\eta}_{t-1}] \\ &+ \mathbb{C}\text{ov}[\tilde{\eta}_t - \tilde{o}_t, Z_{o,t} - Z_{\eta,t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, \tilde{o}_t - \tilde{\eta}_{t-1}] \\ &+ \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, Z_{o,t} - Z_{\eta,t-1}] \\ &= \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, Z_{o,t} - Z_{\eta,t-1}]\,,\end{aligned} \tag{A.13}$$

where the first three terms are zero since the actual returns are uncorrelated and independent from the bid-ask bounces. By expanding the last term we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, Z_{o,t} - Z_{\eta,t-1}] &= \mathbb{C}\text{ov}[Z_{\eta,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[-Z_{o,t}, Z_{o,t}] \\ &+ \mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, -Z_{\eta,t-1}]\,.\end{aligned} \tag{A.14}$$

Since the random variables Z are independent for different trades, we might assume that the only non-vanishing term is $\mathbb{C}\text{ov}[-Z_{o,t}, Z_{o,t}] = -\mathbb{V}[Z]$. However, we should pay extra care when the open price is selected as the high or low price for period t. By decomposing the covariance with Equation (A.1), we have:

$$\begin{aligned}\mathbb{C}\text{ov}[Z_{\eta,t}, Z_{o,t}] &= 0.5(\mathbb{C}\text{ov}[Z_{o,t}, Z_{h,t} + Z_{l,t}]) \\ &= 0.5(\mathbb{C}\text{ov}[Z_{o,t}, Z_{h,t}] + \mathbb{C}\text{ov}[Z_{o,t}, Z_{l,t}]) \\ &= 0.5\mathbb{E}[\mathbb{C}\text{ov}[Z_{o,t}, Z_{h,t} \mid Z_{o,t} = Z_{h,t}]] \\ &\quad + 0.5\mathbb{E}[\mathbb{C}\text{ov}[Z_{o,t}, Z_{l,t} \mid Z_{o,t} = Z_{l,t}]] \\ &= \frac{1}{2}\mathbb{V}[Z_p]\big(\mathbb{P}[Z_{o,t} = Z_{h,t}] + \mathbb{P}[Z_{o,t} = Z_{l,t}]\big)\,.\end{aligned}$$

where $\mathbb{P}[Z_{o,t} = Z_{h,t}]$ and $\mathbb{P}[Z_{o,t} = Z_{l,t}]$ are the probabilities that the open price is selected as the high or low price respectively, and where the variance of Z_p depends on the probability p of the high price to be buyer initiated or, equivalently, of the low price to be seller initiated. The last term left to compute is $\mathbb{C}\text{ov}[Z_{\eta,t} - Z_{o,t}, -Z_{\eta,t-1}]$. This 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) is no trade is observed for period t then $Z_{\eta,t} - Z_{o,t} = 0$.
We estimate the probability that the open price is selected as the high or low price by counting the fraction of times in which the open price matches the high ($\nu_{o=h}$) or low ($\nu_{o=l}$) price:

$$\mathbb{P}[Z_{o,t} = Z_{h,t}] \hat{=} \nu_{o=h}\,, \quad \mathbb{P}[Z_{o,t} = Z_{l,t}] \hat{=} \nu_{o=l}\,.$$

By considering Equation (A.14) and rewriting Equation (A.13), we have:

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

where $k = \mathbb{V}[Z_p]/\mathbb{V}[Z] = 4p(1-p)$ is the ratio between the variance of Z_p with a generic probability p and the variance of Z with $p = \frac{1}{2}$. The final formula is obtained

by computing the variance of Z in Appendix A.1:

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

A.2.7 CHLO prices

By replacing η_t with o_t and following the same steps illustrated in Section A.2.5, we obtain:

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

A.2.8 OHLC prices

By replacing η_{t-1} with c_{t-1} and following the same steps illustrated in Section A.2.6, we obtain:

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

A.3 The Efficient Generalized Estimator

This section provides the optimal way to combine our generalized estimators in Table 1 to minimize the estimation variance and obtain an efficient estimator.

A.3.1 Moment Conditions and GMM

We notice that all the estimators share the common structure:

$$S^2 = -\frac{4\mathbb{C}\text{ov}[r_{1,t}, r_{2,t}]}{\nu} = -4\frac{\mathbb{E}[r_{1,t}r_{2,t}] - \mathbb{E}[r_{1,t}]\mathbb{E}[r_{2,t}]}{\nu} \approx -\frac{4\mathbb{E}[r_{1,t}r_{2,t}]}{\nu}\,, \tag{A.15}$$

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 (A.15), we can rewrite each estimator as a moment condition:

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

Let us introduce, for each estimator i, the random vector $X_{i,t} = -4r_{1,t}^{(i)}r_{2,t}^{(i)}/\nu^{(i)}$ 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:

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

By applying GMM, the efficient estimator is given by:

$$\hat{S}^2 = \arg\min_{S^2} \sum_{ij} (S^2 - \bar{X}_i^\intercal)\Omega_{ij}(S^2 - \bar{X}_j)\,, \tag{A.16}$$

where the weighting matrix is the inverse of the variance-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 (A.16), setting the derivative equal to zero, and solving for S^2, we obtain:

$$\hat{S}^2 = \frac{\sum_i \bar{X}_i \sum_j \Omega_{ij}}{\sum_{ij}\Omega_{ij}} = \sum_i w_i \bar{X}_i \quad \text{with} \quad \begin{cases} \Omega = \mathbb{V}[X_t]^{-1} \\ w_i = \dfrac{\sum_j \Omega_{ij}}{\sum_{ij}\Omega_{ij}} \end{cases}. \tag{A.17}$$

A.3.2 Prior Knowledge

In principle, we could apply GMM using all the estimators in Table 1, that is, eight moment conditions that would lead to an 8×8 covariance matrix. Although the approach is asymptotically efficient, it is expected to perform poorly on small samples due to the

[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.

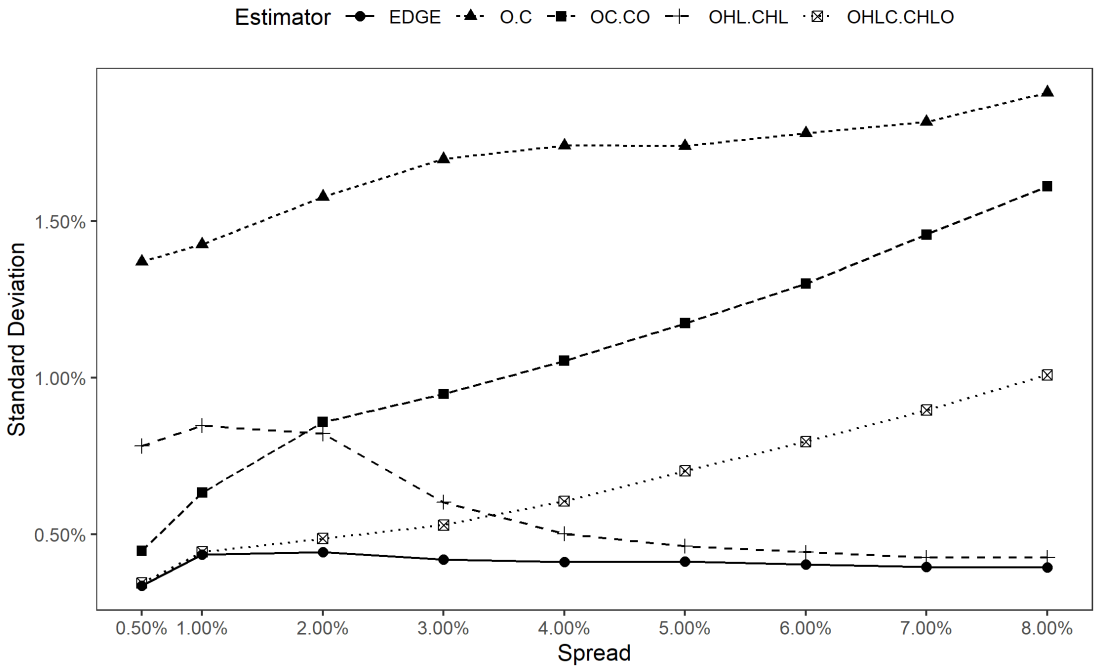


Figure A.1: Comparison of the standard deviation of bid-ask spread estimates based on EDGE and the OHLC estimators, for several spread levels (horizontal axis), as described in Section 2.2.2. All the estimators are unbiased, and the minimum-variance estimator coincides with the best estimator in the usual root mean squared error sense.

noise in the estimation of the large covariance matrix. For this reason, we introduce prior knowledge on the form of the covariance matrix and reduce the number of entries that need to be estimated.

First, we notice that the covariances in the left column of Table 1 are due to the bid-ask spread incorporated in the open (o_t) prices. The covariances in the right column are due instead to the close (c_{t-1}) prices. These estimators are expected to be weakly correlated. Moreover, after dropping all the periods t with no trades such that $h_t = l_t = c_{t-1}$, the estimators are affected by the same variance. Therefore, taking the pairwise average provides an estimator that is superior to both of them. We refer to these estimators as the O-C, OC-CO, OHL-CHL, OHLC-CHLO estimators.

Then, we notice that O-C is dominated by OC-CO and OC-CO is dominated by OHLC-CHLO, while OHL-CHL exhibits a different behaviour, as represented in Figure A.1. The minimum variance estimator is OHLC-CHLO for small spreads and OHL-CHL for large spreads.[22]

[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 big 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 and 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).

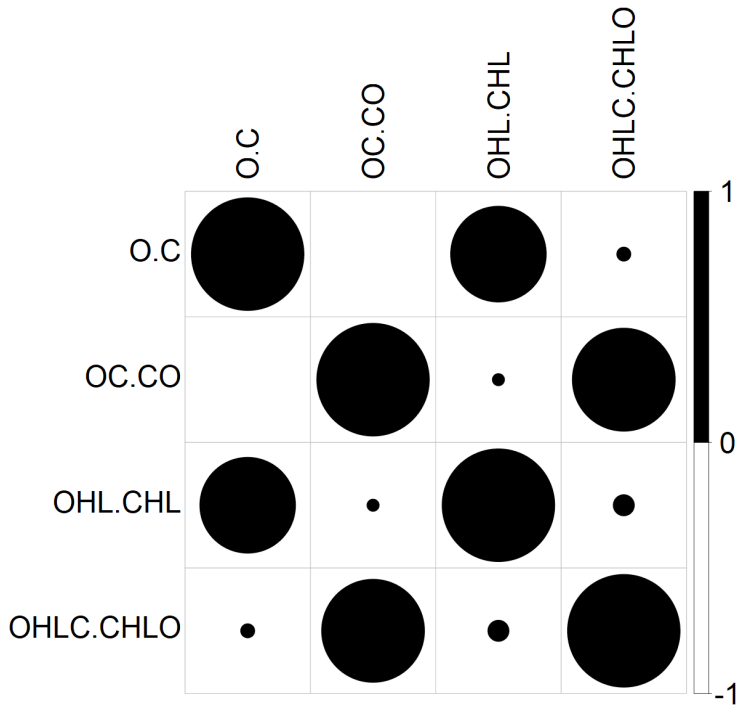


Figure A.2: Correlation matrix between OHLC bid-ask spread estimators, for a simulated price process as described in Section 2.1.1. The simulation uses 390 trades per day with 100% probability of observing a trade and a constant spread of 1%. The size of the circles is proportional to the correlation between the estimators.

The two estimators are summarized below:

$$\text{OHL-CHL}: \quad S^2 = -2\frac{\mathbb{C}\text{ov}[\eta_t - o_t, o_t - \eta_{t-1}] + \mathbb{C}\text{ov}[\eta_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{1 - k\nu_{o,c=h,l}},$$

$$\text{OHLC-CHLO}: \quad S^2 = -2\frac{\mathbb{C}\text{ov}[\eta_t - o_t, o_t - c_{t-1}] + \mathbb{C}\text{ov}[o_t - c_{t-1}, c_{t-1} - \eta_{t-1}]}{1 - k\nu_{o,c=h,l}},$$

where we set the adjustment for infrequent trades equal to the average adjustment $\nu_{o,c=h,l} = (\nu_{o=h,l} + \nu_{c=h,l})/2$. We consider both the estimators as moment conditions so that the GMM covariance matrix reduces to a 2×2 matrix with 3 entries to be estimated: σ_1^2 (variance of OHLC-CHLO), σ_2^2 (variance of OHL-CHL), and σ_{12} (covariance between the two estimators). As OHLC-CHL and OHL-CHL are based on different log-returns with minimal overlap, we expect the two estimators not to be strongly correlated (see Figure A.2). This leads us to set $\sigma_{12} = 0$ so that we are left with a diagonal variance matrix with entries σ_1^2 and σ_2^2. Therefore, according to Equation (A.17), the efficient estimator is given by:

$$S^2 = -2\frac{w_1\mathbb{E}[X_1] + w_2\mathbb{E}[X_2]}{1 - k\nu_{o,c=h,l}}, \tag{A.18}$$

with X_1, X_2 defined in Equation (12) and the weights provided in Equation (13). In the next section, we provide an estimator for k that leads to the final formula in Equation (11).

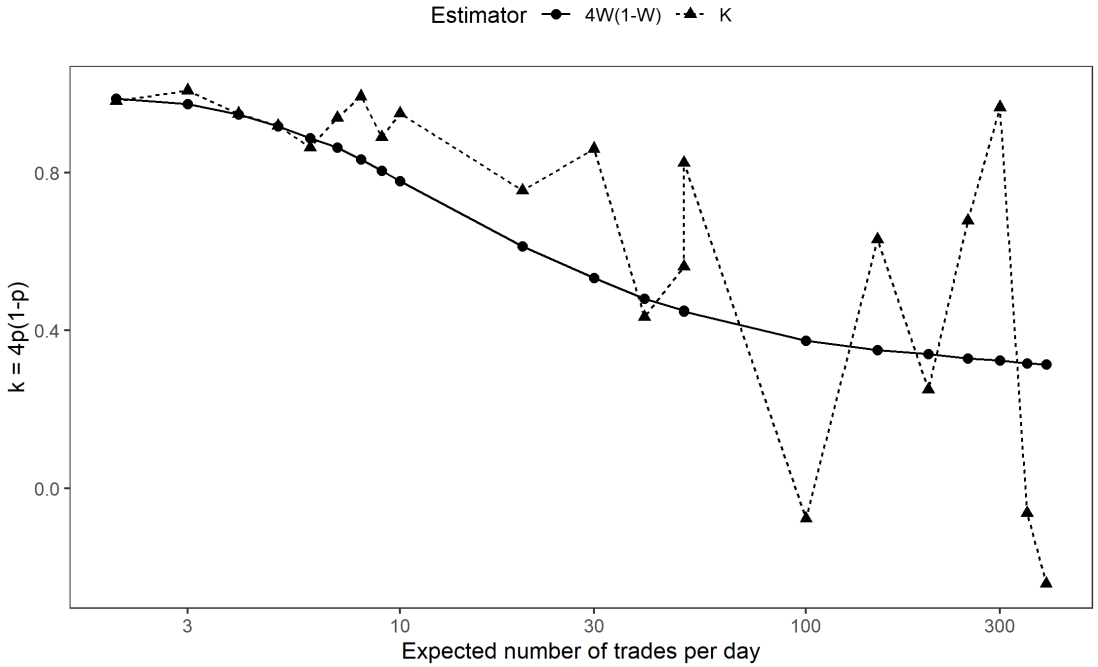


Figure A.3: Comparison between the two estimators for $k = 4p(1-p)$ as described in Section A.3.3, for a simulated price process as described in Section 2.1.1. The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per day is specified in the horizontal axis. The simulations use a constant spread of 1%. The estimator $k = 4w_1w_2$ is smoother (solid line) than the benchmark estimator (dotted line).

A.3.3 Estimation of k

We now need to estimate $k = 4p(1-p)$ where p 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 if the probability p of the high price to be buyer initiated is high (low), then the spread must be big (small) compared to the asset's volatility.[23] In this case, the minimum-variance estimator is OHL-CHL (OHLC-CHLO) as shown in Figure A.1 and the efficient estimator will increase (decrease) the weight w_2. This leads us to identify p with w_2, where w_2 is the weight of the OHL-CHL estimator. Thus, we estimate $k = 4p(1-p) = 4w_2(1-w_2) = 4w_1w_2$ where w_1 is the weight of the OHLC-CHLO estimator.

In Figure A.3, we benchmark this estimator against a naive estimation obtained using Equation (A.18), which depends on k, and the OC-CO estimator, which does not depend on k. This leads to a system of two equations in two unknowns (S^2 and k) that can be easily solved for k. We find the estimator based on w_1 and w_2 to be much smoother and more precise than the benchmark estimator.

Finally, we notice that a precise estimate of k is only needed when the number of trades per period is low. Otherwise the adjustment ν will be close to zero such that the denominator in Equation (A.18) will be close to 1 regardless of the value of k. In other words, we don't expect k to drive the spread estimates, but rather to represent a fine adjustment. By setting $k = 4w_1w_2$ in Equation (A.18), we obtain Equation (11).

[23] 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.

A.4 Further Simulation Results

Table A.1

Estimated Hourly Spreads in High Frequency

Hourly spread estimates from EDGE as proposed in this paper and the ones obtained with the estimators in Abdi and Ranaldo (2017) (AR and AR2), Corwin and Schultz (2012) (CS and CS2), and Roll (1984) for a simulated price process as described in Section 2.1.2. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of spread estimates, and the proportion of spread 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: Simulated Spread Estimates under Near-Ideal Conditions							
Spread 0.10%	Mean	0.10%	0.10%	0.08%	0.08%	0.10%	0.10%
	σ	0.01%	0.02%	0.01%	0.02%	0.01%	0.06%
	$\%\leq 0$	0.00%	0.76%	0.00%	0.00%	0.00%	17.02%
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%	0.28%
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%	0.01%
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: Only 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%
	$\%\leq 0$	5.30%	12.98%	0.00%	67.92%	0.00%	20.79%
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%	0.07%	0.00%	7.80%	0.00%	1.24%
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%	0.09%
	$\%\leq 0$	0.00%	0.00%	0.00%	0.36%	0.00%	0.04%
Spread 1.00%	Mean	0.99%	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%	0.07%	0.00%	0.00%

Table A.2

Estimated Yearly Spreads in Low Frequency

Yearly spread estimates from EDGE as proposed in this paper and the ones obtained with the estimators in Abdi and Ranaldo (2017) (AR and AR2), Corwin and Schultz (2012) (CS and CS2), and Roll (1984) for a simulated price process as described in Section 2.1.1. For each assumed spread level, Panel A reports the mean spread estimate, the standard deviation of spread estimates, and the proportion of spread estimates that are nonpositive across the simulations. Panel B reports results from simulations incorporating overnight returns. In these simulations, overnight returns are normally distributed with mean zero and standard deviation 1.5%.

		EDGE	AR	AR2	CS	CS2	Roll
Panel A: Near-Ideal Conditions							
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%
	$\%\leq 0$	39.87%	50.23%	0.00%	13.53%	0.00%	48.17%
Spread 0.20%	Mean	0.19%	0.37%	1.19%	0.27%	1.27%	0.69%
	σ	0.13%	0.42%	0.10%	0.17%	0.09%	0.75%
	$\%\leq 0$	23.35%	47.69%	0.00%	6.54%	0.00%	46.61%
Spread 0.30%	Mean	0.26%	0.39%	1.19%	0.35%	1.32%	0.68%
	σ	0.15%	0.42%	0.10%	0.17%	0.09%	0.74%
	$\%\leq 0$	13.86%	45.74%	0.00%	2.64%	0.00%	46.46%
Spread 0.40%	Mean	0.38%	0.44%	1.21%	0.44%	1.38%	0.72%
	σ	0.14%	0.43%	0.10%	0.18%	0.09%	0.76%
	$\%\leq 0$	4.21%	40.22%	0.00%	0.83%	0.00%	44.63%
Spread 0.50%	Mean	0.48%	0.50%	1.22%	0.53%	1.44%	0.78%
	σ	0.13%	0.45%	0.10%	0.18%	0.10%	0.78%
	$\%\leq 0$	1.68%	35.56%	0.00%	0.18%	0.00%	40.97%
Panel B: Overnight Returns							
Spread 0.10%	Mean	0.28%	0.40%	1.34%	0.06%	1.18%	0.72%
	σ	0.33%	0.46%	0.11%	0.10%	0.09%	0.83%
	$\%\leq 0$	50.52%	50.08%	0.00%	54.02%	0.00%	48.83%
Spread 0.20%	Mean	0.30%	0.40%	1.35%	0.09%	1.22%	0.71%
	σ	0.33%	0.46%	0.11%	0.12%	0.09%	0.81%
	$\%\leq 0$	47.17%	49.52%	0.00%	41.27%	0.00%	48.44%
Spread 0.30%	Mean	0.33%	0.42%	1.35%	0.13%	1.26%	0.73%
	σ	0.35%	0.47%	0.12%	0.14%	0.09%	0.81%
	$\%\leq 0$	43.69%	47.98%	0.00%	30.68%	0.00%	47.51%
Spread 0.40%	Mean	0.37%	0.46%	1.37%	0.19%	1.31%	0.78%
	σ	0.36%	0.47%	0.11%	0.16%	0.09%	0.83%
	$\%\leq 0$	39.02%	42.72%	0.00%	18.42%	0.00%	44.98%
Spread 0.50%	Mean	0.47%	0.53%	1.38%	0.25%	1.36%	0.82%
	σ	0.37%	0.50%	0.11%	0.17%	0.10%	0.85%
	$\%\leq 0$	29.49%	38.36%	0.00%	9.88%	0.00%	43.43%

Table A.3

Stress Test

Panel A reports the correlation coefficient (Cor.), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE), achieved by the estimators in the simulations described in Figure 5, using a rolling window ranging from one year to one month. Panel B reports the same metrics for simulations performed in high frequency, where we use an estimation window ranging from 10 minutes to one day. These simulations consist of 252 8-hour stock-day, and each day consists of $8 \times 60 \times 60 = 28800$ seconds. For each second, 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 second and x is a random draw from a unit normal distribution. The daily standard deviation equals 3%, and the standard deviation per second equals 3% divided by $\sqrt{28800}$. The bid (ask) for each second is defined as P_m multiplied by one minus (plus) half the assumed bid-ask spread. The probability of observing a trade ranges from 5% to 95% and varies over time according to $p = 0.5 + 0.45 \times \cos(\frac{20\pi t}{n})$ where $t = 1, 2, ...$ represents the time index and $n = 1000 \times 8 \times 60 \times 60$ is the total number of seconds in the simulation. The deterministic component of the spread varies over time according to $\mu = 0.003 \times (1 + \sin(\frac{2\pi t}{n}))$, for an average spread of 0.3%. Then, for each second, the spread is randomly drawn from a normal distribution with mean μ and standard deviation 0.001. Negative spreads are set to zero. For each minute, we use a rolling window of the previous 10 minutes, 1 hour, or 1 day (8×60 minutes) to estimate the spread. The estimates are benchmarked with the average spread in the corresponding window.

		EDGE	AR	AR2	CS	CS2	Roll
Panel A: Low Frequency							
1 Month	Cor.	96.61%	91.28%	90.55%	83.46%	82.43%	62.67%
	MAPE	31.98%	53.05%	50.04%	57.55%	80.77%	105.66%
	RMSE	0.54%	0.85%	0.92%	1.29%	1.35%	2.00%
6 Months	Cor.	99.37%	97.05%	95.24%	86.03%	84.78%	89.39%
	MAPE	21.59%	32.85%	46.62%	51.32%	79.61%	61.71%
	RMSE	0.30%	0.50%	0.78%	1.17%	1.26%	0.96%
12 Months	Cor.	99.67%	97.84%	95.78%	86.33%	85.06%	93.62%
	MAPE	20.18%	27.83%	46.40%	50.80%	79.52%	50.99%
	RMSE	0.27%	0.43%	0.76%	1.15%	1.25%	0.74%
Panel B: High Frequency							
10 Mins	Cor.	93.16%	90.14%	88.13%	80.61%	79.90%	65.36%
	MAPE	29.96%	38.00%	29.96%	41.73%	43.05%	74.91%
	RMSE	0.08%	0.09%	0.10%	0.13%	0.13%	0.19%
1 Hour	Cor.	95.59%	89.36%	85.69%	78.12%	76.88%	89.56%
	MAPE	23.59%	26.38%	25.57%	38.58%	42.87%	42.00%
	RMSE	0.06%	0.09%	0.12%	0.15%	0.14%	0.09%
1 Day	Cor.	99.49%	89.93%	85.26%	77.66%	76.31%	95.21%
	MAPE	21.30%	23.18%	24.81%	38.36%	43.02%	26.08%
	RMSE	0.03%	0.09%	0.12%	0.15%	0.15%	0.06%

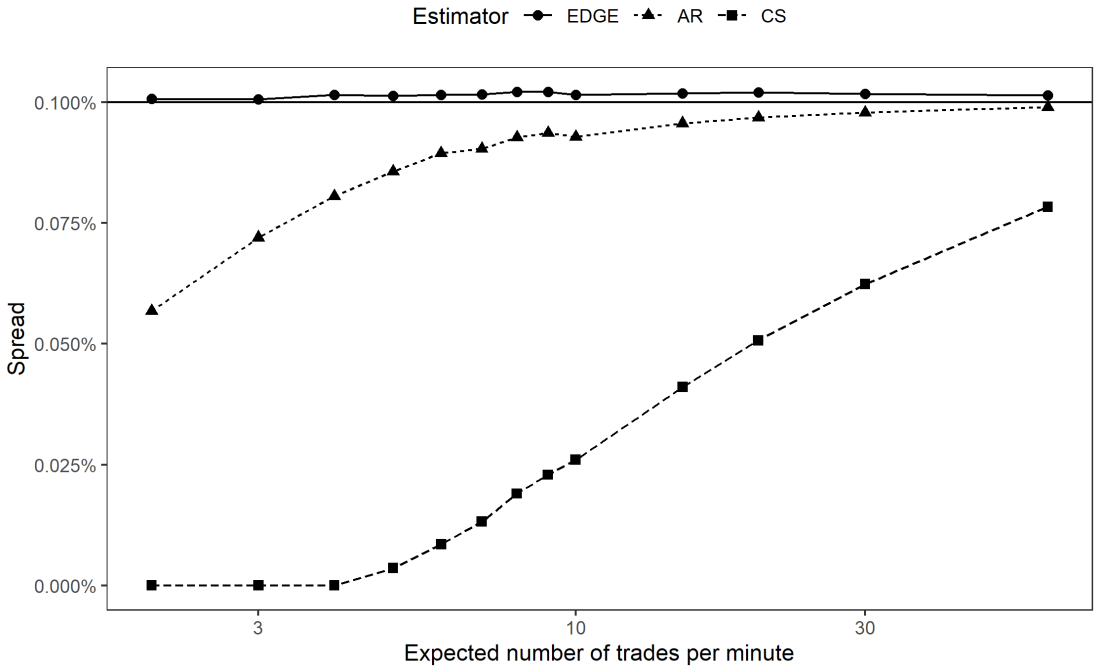


Figure A.4: Comparison of high-frequency spread estimates based on EDGE as proposed in this paper with the estimators by Corwin and Schultz (2012) (CS) and Abdi and Ranaldo (2017) (AR), for a simulated price process as described in Section 2.1.2. The probability of observing a trade ranges from 0.5% to 100% and the corresponding expected number of trades per minute is specified in the horizontal axis. The simulations use a constant spread of 0.10%.

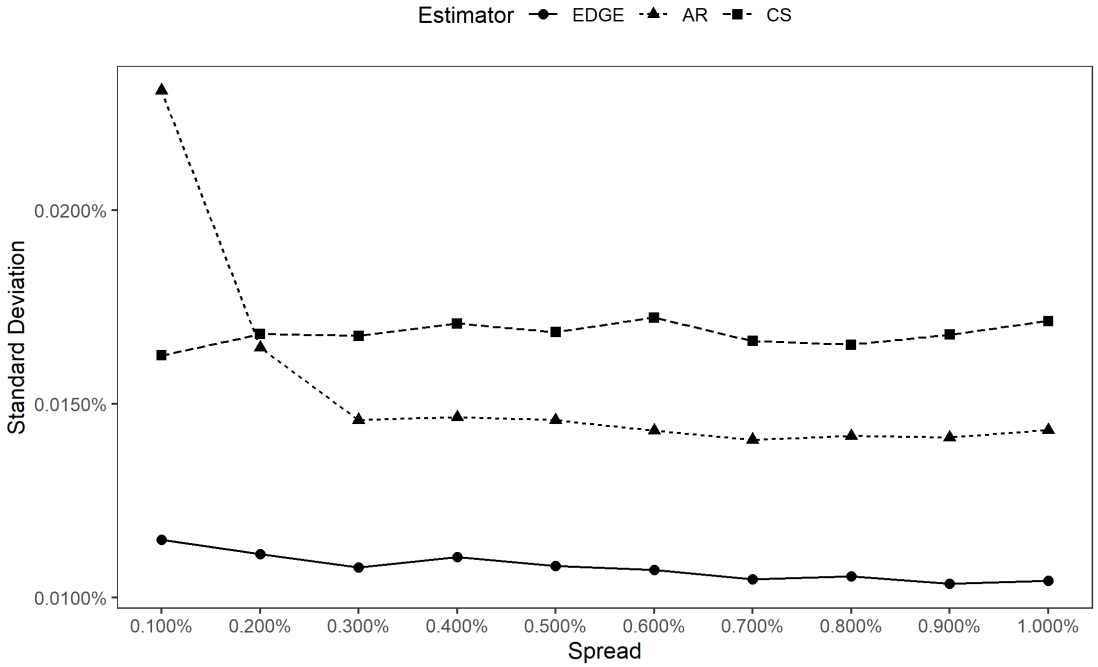
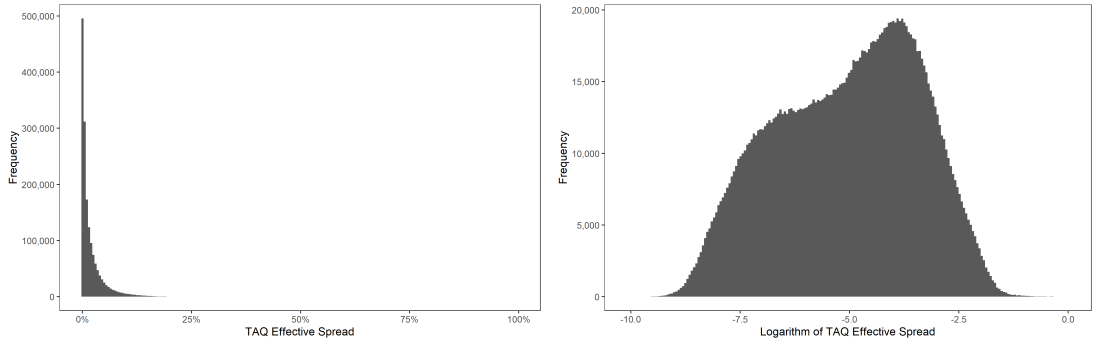


Figure A.5: Comparison of the standard deviation of high-frequency spread estimates based on EDGE as proposed in this paper with the estimators by Corwin and Schultz (2012) (CS) and Abdi and Ranaldo (2017) (AR), for several spread levels (horizontal axis) as described in Section 2.2.2. These simulations use 60 trades per minute.

A.5 Further Empirical Results

The distribution of the TAQ effective spreads is highly skewed as displayed in Figure A.6a. 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 on 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 A.6b. As the argument of the logarithm must be strictly positive, we use only the positive estimates produced by the estimators. Table A.4 shows the MAPE and RMSE computed on the logarithm of the positive estimates. A comparison on the fraction of non-positive estimates is given in Table 4.

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



(a) TAQ Effective Spreads

(b) Logarithm of TAQ Effective Spreads

Figure A.6: The histograms show the empirical distribution of monthly TAQ effective spreads. Figure (a) reports the distribution of the spreads. Figure (b) reports the distribution of the logarithm of the spreads.

Table A.4

MAPE and RMSE with TAQ Effective Spreads

The table shows group specific Mean Absolute Percentage Errors (MAPE) and Root Mean Squared Errors (RMSE) of spread estimates with the TAQ effective spread as described in Section A.5. The lowest MAPE and the lowest RMSE per group are highlighted in bold. EDGE is the estimator proposed in this paper, AR and AR2 are the estimators proposed by Abdi and Ranaldo (2017), CS and CS2 are the estimators proposed by Corwin and Schultz (2012), and the Roll (1984) estimator. All estimators are based on daily observations using a monthly estimation window. The sample period is from 1993–2020 (CRSP-TAQ merged sample).

	MAPE (%)						RMSE					
	EDGE	AR	AR2	CS	CS2	Roll	**EDGE**	AR	AR2	CS	CS2	Roll
Panel A: Analysis across different markets												
NYSE	**21**	25	23	**21**	24	32	1.8	2.0	1.9	**1.6**	1.9	2.5
AMEX	**14**	17	21	50	44	18	**0.8**	**0.8**	1.0	2.7	2.9	1.0
NASDAQ	**16**	19	22	37	37	23	**1.0**	1.2	1.3	2.2	2.5	1.5
Panel B: Analysis across time periods												
1993–1996	**9**	13	18	50	43	14	**0.5**	0.6	0.8	2.9	3.1	0.7
1997–2000	**12**	15	17	42	32	18	**0.6**	0.8	0.9	2.3	2.4	1.0
2001–2002	**15**	18	20	37	35	23	**0.9**	1.1	1.1	2.1	2.4	1.4
2003–2007	**20**	22	22	25	28	27	**1.4**	1.6	1.6	1.6	2.0	2.0
2008–2011	**28**	32	32	33	40	38	**1.7**	1.9	1.9	**1.7**	2.1	2.4
2012–2015	**21**	24	24	22	28	30	1.6	1.8	1.8	**1.5**	2.0	2.3
2016–2020	22	25	25	**21**	27	33	1.7	2.0	1.9	**1.5**	2.0	2.5
Panel C: Analysis across market capitalization												
Quintile 1	**17**	20	25	55	52	22	**0.7**	0.8	1.0	2.7	3.1	1.0
Quintile 2	**13**	17	21	46	45	18	**0.8**	0.9	1.1	2.6	3.0	1.0
Quintile 3	**14**	16	18	31	28	20	**0.9**	1.0	1.1	1.9	2.0	1.3
Quintile 4	**18**	21	21	21	22	28	**1.4**	1.6	1.5	1.5	1.7	2.0
Quintile 5	24	28	26	**20**	26	35	2.0	2.2	2.1	**1.6**	2.1	2.7
Panel D: Analysis across spread sizes												
Quintile 1	26	30	29	**19**	28	38	2.1	2.4	2.3	**1.6**	2.2	2.9
Quintile 2	20	24	22	**18**	21	31	1.5	1.8	1.6	**1.4**	1.6	2.2
Quintile 3	**15**	18	17	22	20	24	**1.1**	1.2	1.2	1.4	1.5	1.6
Quintile 4	**12**	14	16	34	29	17	**0.7**	0.8	0.9	2.0	2.0	1.0
Quintile 5	**15**	19	29	70	71	19	**0.5**	0.7	1.0	3.2	3.8	0.7

Table A.5

Correlation with Yearly TAQ Effective Spreads

The table shows group specific correlations of spread estimates with the TAQ effective spread. The table also reports the median effective spread per group and the fraction of spread estimates that are non-positive. 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, AR and AR2 are the estimators proposed by Abdi and Ranaldo (2017), CS and CS2 are the estimators proposed by Corwin and Schultz (2012), and the Roll (1984) estimator. All estimators are based on daily observations using a yearly estimation window. The sample period is from 1993–2020 (CRSP-TAQ merged sample).

		Correlation (%)						$\%\leq 0$			
Group	Spread	**EDGE**	AR	AR2	CS	CS2	Roll	**EDGE**	AR	CS	Roll
Panel A: Analysis across different markets											
NYSE	0.17%	**55**	46	49	44	41	9	**35**	41	36	44
AMEX	1.90%	**78**	67	66	48	51	4	**20**	25	41	34
NASDAQ	1.47%	**85**	77	69	46	42	25	**10**	17	**10**	26
Panel B: Analysis across time periods											
1993–1996	2.68%	**88**	81	71	48	47	43	**11**	17	22	25
1997–2000	1.82%	**83**	78	70	48	47	52	**16**	25	30	33
2001–2002	1.42%	**85**	82	77	58	58	54	**17**	24	29	33
2003–2007	0.35%	**77**	67	69	45	49	6	19	25	**18**	35
2008–2011	0.29%	**77**	64	63	40	38	6	16	21	**13**	27
2012–2015	0.20%	**71**	64	62	42	36	10	23	31	**9**	38
2016–2020	0.20%	**65**	53	54	49	40	16	28	31	**13**	34
Panel C: Analysis across market capitalization											
Quintile 1	3.54%	**80**	73	65	42	40	25	**9**	15	16	24
Quintile 2	2.22%	**79**	67	53	35	23	7	**10**	16	17	25
Quintile 3	1.16%	**82**	63	54	37	26	11	**14**	20	17	30
Quintile 4	0.32%	**83**	67	61	53	42	17	24	31	**21**	37
Quintile 5	0.09%	**42**	33	34	34	28	1	30	39	**27**	41
Panel D: Analysis across spread sizes											
Quintile 1	0.09%	22	15	**27**	16	23	0	34	41	**27**	43
Quintile 2	0.27%	**51**	34	44	40	37	1	28	35	**25**	41
Quintile 3	0.82%	**67**	49	51	48	39	3	**18**	26	20	37
Quintile 4	1.94%	**75**	62	57	41	34	10	**9**	16	15	27
Quintile 5	4.71%	**75**	68	58	37	36	24	**4**	7	13	13