A problem of zeros: The misbehavior of simple bid-ask spread estimators*

Pedro Tremacoldi-Rossi† Scott H. Irwin†‡

February 19, 2019

Abstract

We show that widely used bid-ask spread estimators based on high, low, and closing prices are inadequate proxies when markets are very liquid. For a given volatility level and maintaining all model assumptions, we demonstrate how smaller true spreads induce these measures to yield negative estimates more often and positive estimates with larger upward bias. Using a simple framework, we establish how the frequency of negative spreads identifies the bias empirically. Effective spreads in many modern markets result in poor performance of the high-low and close-high-low estimators, which cannot be ameliorated with any adjustment proposed in the literature.

Keywords: bid-ask spread, estimation bias, liquidity costs

JEL Codes: C10, G12, G13

*We are grateful to Adam Clark-Joseph, Phil Garcia, Tatiana Mocanu, Esen Onur, Michel Robe, Teresa Serra, and seminar participants at the the University of Illinois for comments. Quanbiao Shang provided assistance with data. Any errors are our own.

†Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign.

‡Corresponding author: sirwin@illinois.edu.

1 Introduction

The explosion in high frequency data availability over the last decade enabled the direct computation of bid-ask spreads (BAS), thereby easing the measurement of trading costs in financial markets. Nonetheless, there is still substantial interest in "simple" bid-ask spread estimators that rely on low frequency data. The reasons for developing and using simple spread proxies are severalfold. First, obtaining trade-and-quote (TAQ) data is costly and some times infeasible, either because of market structure (Schestag et al. (2016)) or when the sample period precedes electronic trading (Hasbrouck (2009)). Second, simple spread estimators can greatly simplify the burdensome task of constructing daily spreads from intraday data in studies with other variables available at a daily frequency.

The availability of TAQ data allows performance of simple estimators to be tested against the "true" spread. These horserace-type exercises inform which proxies should perform well whenever a practitioner needs to estimate bid-ask spreads.[1] Although richer trading data improved the assessment of spread estimator performance, understanding of why some estimators are better than others is still limited. More importantly, even for estimators that perform well on average, empirical issues are pervasive, and their implications for the quality of proxies remains largely unexplained.

Consider the shortcoming of returning negative or indeterminate estimates that is often displayed by spread estimators. Even two of the best performing simple bid-ask spread estimators, the high-low estimator (HL) (Corwin and Schultz (2012)), and the close-high-low estimator (CHL) (Abdi and Ranaldo (2017)) are negative about half the time for any asset that is reasonably liquid, from live cattle futures to Apple stocks. The common *ad hoc* remedy of setting negative estimates to zero (Goyenko et al. (2009), Hasbrouck (2009), and Karnaukh et al. (2015)) reflects the predominant, yet untested, view that negative spread estimates represent values that are missing "at random", or that arise as a result of the violation of model assumptions. Empirically, the adjustment leads to an overlooked consequence: zero imputation tends to produce average HL and CHL estimates that are more accurate than averages taken over positive-only estimate values. Why would zero imputation improve an estimator's performance? If positive estimates alone are

[1] For examples, see Goyenko et al. (2009) and Fong et al. (2017). Of course, out-of-sample validity of good proxies depends on the tacit assumption that the in-sample accuracy of an estimator extends to different settings. Since nothing guarantees that will be the case, good BAS proxies may simply fail to measure trading costs when applied to "untested" data (e.g. Locke and Venkatesh (1997)).

accurate, the sample average would become downward-biased after zero imputation. Furthermore, given their empirical recurrence, could negative spreads be informative of the underlying behavior of spread estimators?

In this paper, we show that the frequency of empirical negative spreads identifies estimation bias of simple bid-ask spread estimators, therefore revealing observable proxy misbehavior of the HL and CHL measures. For a given volatility level and maintaining all model assumptions, smaller spreads induce negative estimates more often and positive estimates with larger upward bias. As a consequence, simple spread estimators suffer from large bias when markets are very liquid, which should constitute nearly ideal empirical settings for spread proxies.

We choose the high-low measure to represent the class of simple spread estimators due to its extensive use in applied work, ranging from commodity (Adams and Glück (2015)) to frontier markets (Marshall et al. (2015)), and from microstructure (Easley et al. (2016)) to how financial research affects returns (McLean and Pontiff (2016)), and because the estimator outperforms most alternative proxies. The close-high-low estimator relaxes some of HL's assumptions, and is also one of the simple bid-ask spread estimators to perform best in evaluation studies.

We start our analysis of the behavior of BAS estimators with standard simulation exercises used in the literature but extended to cover spreads smaller than 0.5%. This region of spread size is crucial to provide performance tests with empirical relevance. At least half of all US stocks have average effective spreads of 0.5% or less (Brogaard et al. (2017), Abdi and Ranaldo (2017)). Futures markets trade with liquidity costs even lower. Marshall et al. (2011) document 24 commodity futures including metals, energy, and agricultural products with an average median effective spread below 0.2%. Highly liquid financial futures such as the E-mini S&P 500 display spreads as low as 0.01% (Clark-Joseph (2013)). Even for less capitalized stock markets, an effective spread of 0.5% represents an accurate benchmark. Fong et al. (2017) compute a median effective spread for 42 stock exchanges around the world of 0.7%. Bond markets trade with spreads below 0.2% on average (Chakravarty and Sarkar (2003)).

Our simulation results show that, even when all model assumptions hold, two major regularities appear when the *ex ante* spread falls below 1%: the bias in the average estimated spread becomes very large; and negative estimates are much more frequent, amounting to at least 40% of observations when the latent spread is smaller than 0.5%. Since the proportion of negative estimates is empirically observable, while the true spread and, as a consequence, bias, are unobserved, we explore whether there is information in the frequency of negative estimated spreads that might

identify the bias empirically.

In order to isolate the determinants of proxy misbehavior, we employ a simple framework for each spread estimator that determines when estimates are negative. These negativity conditions are obtained directly from each estimator's functional form and do not depend on violations of model assumptions. Driven by both the true spread and volatility levels, the negativity conditions are more frequently attained when the spread decreases and volatility increases. We show that the spread size accounts for most of the observable negative spreads, since shocks to the latent spread are relatively more important than volatility shocks.

We then connect this finding to bias by showing that the same mechanism inducing negative spreads results in the misbehavior of simple spread proxies. As the true spread narrows, the upward bias of the positive-only estimates increases, regardless of sample size. Imputing zeros does reduce bias marginally, albeit the resulting overstated proxy can be as high as 10 times the true spread value. To illustrate how our framework works in practice, we apply the HL and CHL measures to a large commodity futures market, specifically, the corn futures market. With an effective spread of 0.1%, corn futures are representative of nearly a third of US stocks and all reasonably liquid futures markets. We find strong empirical evidence of misbehavior associated with a large frequency of negative estimates and small effective spreads, as predicted by our framework. This widespread misbehavior goes unnoticed in performance studies based on average measures because their usual setting contains a sizable fraction of illiquid assets. When the spread is large (e.g. more than 1%), the HL and CHL measures tend to perform better, thus driving high average cross-sectional correlations between the proxies and the effective spread.

Our work contributes to the literature in two main ways. First, we uncover a systematic association between the true spread size S_i and bias $\delta_i = S_i - \widehat{Spread}_i$, that might introduce non-classical measurement error in empirical work that uses spread proxies. In a regression model $y_i = \alpha + \beta \widehat{Spread}_i + \varepsilon_i$, assets i with small effective spreads will have larger measurement error than illiquid assets, since spread estimates $\widehat{Spread}_i$ will be more severely biased for smaller unobserved spreads. Finally, by using the observable frequency of negative spread estimates to identify the bias from positive estimates, our findings have concrete applications for empirical work. We suggest practitioners using the HL and CHL measures follow this rule-of-thumb: if the proportion of daily negative spreads is greater than 40%, the spread estimates are severely upward-biased, and the estimators are inadequate proxies for the underlying spread. More generally, simple bid-ask estimators should be used with caution in markets with spreads between 0.5% and 1%, and avoided

altogether in markets with spreads below 0.5%, which corresponds to the majority of stocks and futures markets.

This paper relates to a long tradition of studies focusing on the determinants of misbehavior of trading liquidity proxies, starting with Harris (1990) and more recently with Lou and Shu (2017). Our work is also related to papers that document poor performance of the high-low estimator in specific contexts using simulation exercises (Lin (2014)), and in less than ideal empirical settings (Bleaney and Li (2015) and Nieto (2018)). An important distinction between our results and previous studies on HL is that by focusing on the underlying spread, we are able to provide a framework that identifies proxy misbehavior for assets trading in the most ideal environments, and that is particularly relevant to modern financial markets.[2] This implies that we can also analyze other simple estimators, such as the close-high-low proxy, using the same framework.

The paper is organized as follows: Section 2 introduces the common theoretical framework that simple bid-ask spread estimators rely on. In Section 3, we show simulation results that indicate clear patterns of comovement between the frequency of negative estimates, true spread size and bias in the average HL and CHL estimators. We then demonstrate why spread estimates arise as as consequence of changes in the true spread and volatility levels. In Section 4, we explore the relationship between negative estimates and bias, and argue that the spread size is the most likely driver of bias, therefore allowing bias identification via frequency of negative spreads. We finalize with an empirical application to corn futures markets in Section 5, where we find support for our framework predictions and illustrate its usefulness.

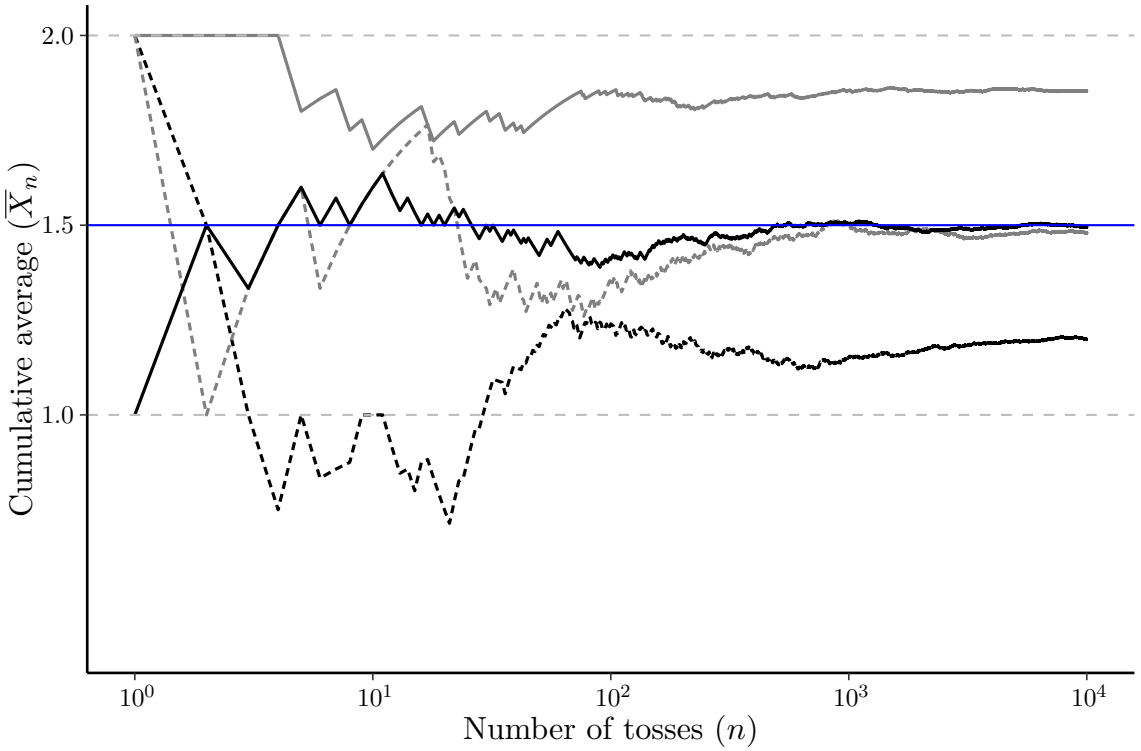
2 Framework

2.1 A conceptual example of misbehavior

Before formally starting our analysis, we provide a conceptual illustration of why obtaining better average estimates with zero imputation might convey information about the estimator's behavior. Simple bid-ask spread estimates are obtained at a daily frequency and usually aggregated to monthly average spreads. Out of the three main *ad hoc* adjustments to deal with negatives - discard the negatives, zero imputation or compute averages over the entire sample - zero imputation

[2] Although our results are not directly conditioned on market microstructure or model assumption violations, modern microstructure might still largely account for very small bid-ask spreads (O'Hara (2015)).

is not only the most widely used adjustment, but arguably increases the accuracy of average spread proxies (Corwin and Schultz (2012)).



Notes: All lines plotted correspond to cumulative averages $\overline{X}_n$ of n independent tosses of a coin with heads = 1 and tails = 2. The black lines are obtained with a fair coin experiment, for which the expected outcome is 1.5 (blue line). The solid black line computes the cumulative average over the entire sequence of tosses, while the dashed black line has zeros imputed in $k \times n$ missing outcomes from the n tosses. The gray lines come from repeating the same experiment with a biased coin, where the probability of obtaining tails in a given toss is 85%. Again, the solid line computes the cumulative sample average over all n tosses, while the dashed line includes imputed zeros. We set the fraction of missing values to $k = 0.2$.

FIGURE I
The Information in Zero Imputation

Say you toss a fair coin n times and assign to heads a value of 1 and to tails a value of 2. The cumulative average of outcomes obtained up to the n-th toss is given by $\overline{X}_n$. In Figure (I), the solid black line shows how the cumulative average $\overline{X}_n$ behaves as n increases. As expected, $\overline{X}_n$ converges to the expected outcome of 1.5 after repeating the toss sufficient times. Imagine that when recording the outcomes of n tosses, the experimenter lost a fraction k of records, with $k \in (0, 1)$. That is, if $k = 0.2$, 20% of n tosses are now missing values. Of course, the cumulative average over the non-missing values, $\overline{X}_{(1-k)n}$, eventually converges to 1.5 as we compensate the sample size loss by tossing the coin more than the initial n times.

Now, what would happen to the cumulative average if we were to replace the $k \times n$ missing values with zeros? The dashed black line shows the cumulative average over $(1-k)n$ outcomes and $k \times n$ zeros. Because zeros weight down the cumulative sample mean, the imputation produces an asymptotic deviation between the true mean of 1.5 and the average over the positive outcomes and zeros, $\overline{X}_{(1-k)n,0}$. What if we replicate the experiment with a biased coin, so that the chance of obtaining tails is 85%, but without informing the experimenter? Because the coin is upward-biased, the cumulative average remains well above 1.5 even with many tosses (solid gray line). However, in the hypothetical of replacing $k \times n$ missing toss outcomes with zero, the cumulative mean over $(1-k)n$ biased outcomes and the zeros (dashed gray line) appears fairly close to the sample average in the unbiased case, and thus to 1.5.

Given that the experimenter is unaware of the coin bias in the second experiment, but knows that the expected outcome is 1.5, an average close to the true mean given by a sample with zeros imputed suggests that the coin must be upward biased. Otherwise, the cumulative average including zeros should replicate the behavior of the black dashed line in the figure. In practice, daily estimates from simple bid-ask spread measures return large values of k. This implies that the adjustment chosen to deal with these "missing" values will affect sample averages. Because negative estimates have no economic meaning, replacing them with zeros does not cause information loss *per se*. The problem of zeros arises from the fact that the adjustment produces more accurate average spreads than using only positive estimates.

In the remainder of this section, we introduce the theoretical framework shared by simple bid-ask spread estimators and provide an empirical illustration of spread proxy misbehavior. The example suggests that the underlying effective spread magnitude, not violations of model assumptions, account for the poor proxy performance.

2.2 The structure of simple spread estimators

A common feature shared across the class of proxies we consider as "simple" BAS estimators is the reliance on easily obtainable data. Another characteristic is that many assumptions embedded in these estimators were first formally used in the popular Roll measure. The convenient dependence on closing price returns and easy computation of Roll's seminal spread estimator (Roll (1984)) have not only made it widely applied, but also a benchmark for the development and comparison of newer simple estimators. Two prominent recent examples are the high-low estimator (Corwin and Schultz

(2012)) and the closely related close-high-low measure (Abdi and Ranaldo (2017)). The motivation of these two measures is to ameliorate some of the deficiencies encountered with the Roll estimator in empirical work, while maintaining desirable characteristics for implementation. One of these deficiencies is Roll's constant empirical indeterminate spread estimates, as we discuss below.

Naturally, many other spread estimators that require only low frequency data for implementation have been proposed. To some extent, most alternative spread proxies that followed or extended Roll (1984) drew information from the same set - daily price returns (French and Roll (1986), Thompson and Waller (1987), Lesmond et al. (1999), Hasbrouck (2004), and Fong et al. (2017)). Because the high-low measure from Corwin and Schultz (2012) introduced a novel identification strategy for the spread, we select it as a proxy to represent simple BAS estimators. As the CHL proxy claims to improve on HL similarly to Garman and Klass (1980) on Parkinson (1980) with the introduction of daily closing prices to the range, we also include the measure in our analysis.

The derivation of simple bid-ask spread estimators begins with the usual assumption that log prices follow a geometric Brownian motion (GBM). Hence, continuous existence of prices during trading hours is implicitly assumed, although continuous observation thereof is not necessary. The relationship between observed (p_t) and true prices ($\mathcal{P}_t$) is given by

$$p_t = \mathcal{P}_t + \frac{S}{2} q_t \tag{1}$$

where $\mathcal{P}_t$ is the latent price, S is the total effective spread and q_t is the order flow indicator ($q_t = 1$ for buyer-initiated orders and $q_t = -1$ for seller-initiated orders). Given that the latent price evolves as a random walk, $\Delta\mathcal{P}_t = \varepsilon_t$, with $\Delta\mathcal{P}_t = \mathcal{P}_t - \mathcal{P}_{t-1}$, $\varepsilon_t \sim \text{NID}\left(0, \sigma^2\right)$, trade direction is assumed to be independent of efficient prices. In spite of its simplicity, (1) represents the cornerstone expression shared by all spread proxies we characterize as simple estimators. Within this category, time is indexed by $t = 1, ..., T$ days, so that only end-of-day low frequency data is necessary to carry out empirical implementation. Daily spread estimates may be averaged at the desired frequency, say M months, in order to compute average spread estimates.

In his seminal work, Roll (1984) only requires data on daily closing prices c_t, so that the relevant realization of observed prices is $c_t = \mathcal{C}_t + q_t S/2$. The "seductive simplicity" of Roll's measure (Harris (1990)) stems largely from requiring only estimates of first-order serial covariance of consecutive daily returns. This enabled researchers and practitioners to easily and cheaply

compute bid-ask spreads, which had largely been unobserved since Demsetz (1968) and others first uncovered their central role in trading costs.

A recurrent empirical indeterminacy of the Roll estimator has received attention of many authors over the years.[3] Harris (1990) was the first to provide a comprehensive account of the mechanisms inducing the misbehavior of the Roll measure, and to suggest *ad hoc* modifications to deal with indeterminate estimates.[4] As we stressed before, zero imputation became a standard practice whenever the estimator produces a negative estimate. The same procedure of imputing zeros is suggested when implementing HL and CHL, where the former can be negative in practice and the latter may have a negative term inside a square root, in the same fashion as the Roll proxy.[5] Zero imputation is argued to improve average spread estimates, and works better than treating negative spreads as missing or averaging negative and positive estimates together.

2.2.1 The high-low estimator

Corwin and Schultz (2012) exploit the empirical regularity that the sign of the trade indicator q_t can be determined at the daily extrema of prices. Under the assumption that high and low prices are buyer- and seller-initiated, respectively, (1) translates as the pair,

$$(h_t, l_t) = \left(\mathcal{H}_t + \frac{S}{2}, \mathcal{L}_t - \frac{S}{2} \right) \tag{2}$$

and more specifically, the daily log range $r_t \equiv h_t - l_t$ can be manipulated by using the moments derived in Parkinson (1980) and Garman and Klass (1980) for the true range $\mathcal{R}$:

$$E\left[T^{-1} \sum_{t=1}^{T} \mathcal{R}_t \right] = k_1 \sigma \quad \text{and} \quad E\left[T^{-1} \sum_{t=1}^{T} \mathcal{R}_t^2 \right] = k_2 \sigma^2 \tag{3}$$

where $k_1 \equiv \sqrt{8/\pi}$ and $k_2 \equiv 4 \ln 2$, and σ^2 is an unbiased (under the assumption of no drift in the GBM) estimator of daily variance based on the range when $E\left[T^{-1} \sum_{t=1}^{T} \mathcal{R}_t \right]$ and $E\left[T^{-1} \sum_{t=1}^{T} \mathcal{R}_t^2 \right]$

[3] Since the estimator is only defined for negative serial covariance, spreads cannot be estimated for a significant portion of financial data that generates positive autocovariance. Moreover, the magnitude of spread estimates depends on the chosen observation interval (daily, weekly etc.), which creates an unnecessary choice for empirical purposes.

[4] More recently, Chen et al. (2017) propose an interesting semiparametric approach to identify the spread in a general framework with flavors of Roll and extended Roll models.

[5] For simplicity, whenever we refer to negative estimates in the context of Roll and CHL estimators, what we mean is that the term evaluated inside the square root in each estimator is negative, and the spread estimate is thus indeterminate.

are replaced with $\mathcal{R}_t$ and $\mathcal{R}_t^2$, respectively.

The Parkinson-Garman-Klauss framework provides the fundamental rationale for the post-Roll generation of simple bid-ask spread estimators. The appealing argument that using additional information contained in the daily range improves volatility estimation, underlies the innovative path taken by Corwin and Schultz (2012). Since the range represents the boundaries of observed daily price oscillation, i.e. the volatility measured in σ^2, it also includes the effective spread at those boundary points according to (2). The additional assumption that the spread is constant over every pair of consecutive days, $\bigcup_{k=1}^{T-1}[t_k, t_{k+1}]$, while volatility remains proportional to time, allows for uniquely determining S from the following expressions:

$$E\left[\sum_{t=1}^{2} r_t^2\right] = (8\ln 2)\,\sigma^2 + \left(8\sqrt{\frac{2}{\pi}}\right)\sigma S + 2S^2 \tag{4}$$

$$E\left[r_t^{*2}\right] = (8\ln 2)\,\sigma^2 + \left(8\sqrt{\frac{1}{\pi}}\right)\sigma S + S^2 \tag{5}$$

which relate the daily range (r_t) and the two-day range (r_t^*), where $r_t^* \equiv \max\{h_t, h_{t+1}\} - \min\{l_t, l_{t+1}\}$, to both the spread and daily volatility. The two-day range captures the overall volatility in each pair of days, and if time proportionality holds, it should correspond to twice the volatility of a single day. After solving for σ^2 in (4) and (5), the high-low spread estimator is obtained as:

$$S = 2\tanh\frac{\alpha}{2} \tag{HL}$$

for

$$\alpha \equiv \Theta\left(\sqrt{\beta} - \sqrt{\gamma}\right), \quad \beta \equiv \sum_{t=1}^{2} r_t^2, \quad \gamma \equiv r_t^{*2} \tag{6}$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, and $\Theta = 1 + \sqrt{2}$. We transform the functional forms for S and α from Corwin and Schultz (2012) for ease of exposition.

We formally define the daily and average forms of the high-low spread estimator below.

DEFINITION 1. *The point-estimate of the high-low spread estimator* (HL) *is simply* $\mathrm{HL}_t \equiv 2\tanh(\alpha/2)$, *for day* t. *The average high-low estimate is* $\overline{\mathrm{HL}} \equiv J^{-1}\sum_{j=1}^{J}\mathrm{HL}^j$, *where* HL^j *is the monthly average high-low estimate for months* $j = 1, ..., J$, $\mathrm{HL}^j = T^{-1}\sum_{t=1}^{T}\mathrm{HL}_t$.

To make Definition 1 concrete, we provide a simple example. After calculating daily-level HL estimates HL_t, we can take averages within each month j, and compute the average high-low estimate over the reference J months. Alternatively, we could simply compute daily estimates HL_t and average over $J = T$ days to obtain the average high-low estimate $\overline{\text{HL}}$.

2.2.2 The close-high-low estimator

The close-high-low spread estimator from Abdi and Ranaldo (2017) combines the use of high and low prices from the HL measure with daily closing prices from Roll (1984). In a sense, the measure uses the Parkinson-Garman-Klauss framework more broadly by integrating the range and daily returns to estimate bid-ask spreads along the lines of HL. The intuition for augmenting the information set is that closing prices are more "contaminated" by the bid-ask bounce than the range (Alizadeh et al. (2002)). For simplicity, consider (2) as a maintained assumption.[6] Let the mid-range, or average price on day t, be defined as

$$\eta_t \equiv \frac{h_t + l_t}{2} = \frac{r_t}{2} + l_t \tag{7}$$

which clearly coincides with the mid-range of daily efficient extreme prices, $\eta_t = (\mathcal{H}_t + \mathcal{L}_t)/2$. A crucial result in Abdi and Ranaldo (2017) establishes how the variance of the mid-range returns relates to the efficient price volatility σ^2:

$$E\left[(\eta_{t+1} - \eta_t)^2\right] = \left(2 - \frac{k_2}{2}\right)\sigma^2 \tag{8}$$

where the variance may be replaced with $\widehat{\sigma}^2_\eta$ for estimated squared returns of consecutive-day mid-range prices. Under the validity of (8), and because the average of consecutive mid-ranges, $\bar{\eta} \equiv (\eta_t + \eta_{t+1})/2$, is an unbiased estimator of the end-of-day midquote, closing prices can be connected to $\bar{\eta}$ so that the CHL spread estimator follows:

$$S = 2\sqrt{(c_t - \eta_t)(c_t - \eta_{t+1})}. \tag{CHL}$$

[6] In reality, the CHL estimator does not depend upon assuming trade direction for extreme daily prices, as in the HL measure. Nonetheless, as Abdi and Ranaldo (2017) recognize, the assumption is generally supported empirically.

Analogous to the Roll estimator, the close-high-low BAS measure is indeterminate if the term $(c_t - \eta_t)(c_t - \eta_{t+1})$ is negative. Similar to before, we distinguish between point-estimates of CHL and average-estimates of the spread proxy in Definition 2:

DEFINITION 2. *The point-estimate of the close-high-low spread estimator* (CHL) *is simply* $\text{CHL}_t \equiv 2\sqrt{(c_t - \eta_t)(c_t - \eta_{t+1})}$, *for day* t. *The average close-high-low estimate is* $\overline{\text{CHL}} \equiv J^{-1}\sum_{j=1}^{J}\text{CHL}^j$, *where* CHL^j *is the monthly average close-high-low estimate for months* $j = 1, ..., J$, $\text{CHL}^j = T^{-1}\sum_{t=1}^{T}\text{CHL}_t$.

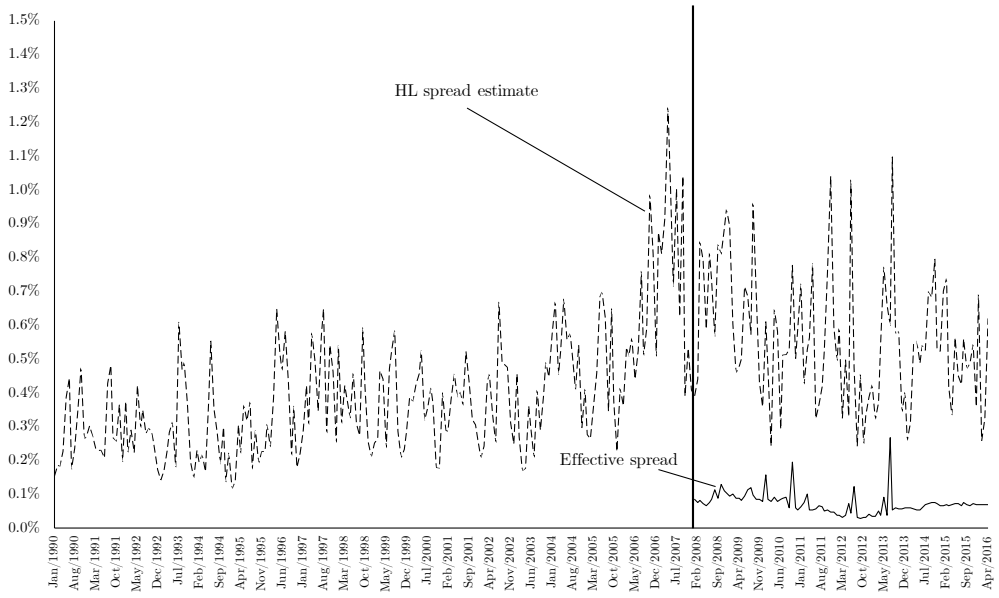
2.3 Related literature on the performance of HL and CHL

The high-low spread estimator has been widely applied since its introduction. An appealing virtue of HL is to use a well-established method in the finance literature to estimate volatility with the log range "backwards". The range suffers from microstructure noise, so after filtering it from volatility effects it is possible to retrieve the bid-ask bounce. As observed by Corwin and Schultz (2012), the HL proxy may still capture transitory volatility, specially in thin markets where large orders can be executed at daily high or low prices. In the case of CHL, mid-ranges do not contain information on S, so identification occurs through closing prices. The only study to directly compare HL and CHL to date is Abdi and Ranaldo (2017). Overall, cross-sectional and time-series correlations, and prediction errors favor the close-high-low over HL as a better spread proxy in stock data.

The high-low estimator was also shown to dominate other low frequency BAS estimators, with equity markets providing in-sample effective spreads and daily prices for evaluation (Corwin and Schultz (2012), Marshall et al. (2015), and Fong et al. (2017)). As mentioned earlier, the average effective spread size in these data fluctuates around 1%. There is some sparse evidence in Corwin and Schultz (2012) and Abdi and Ranaldo (2017) suggesting that bias in HL and CHL estimates is greater for stocks with lower liquidity cost levels.

2.4 An empirical example of misbehavior

Now, we bring both HL and CHL to data to illustrate one example of poor spread proxy performance that is unlikely to be driven by model assumption violations (Figure (II)). The dashed line represents bid-ask spreads for corn futures prices estimated using HL. CHL estimates give nearly identical results. The trend of high-low estimates suggests that trading costs in 2006-2016 increased compared to pre-2000, which contrasts with widespread evidence of systematic decreases in trading costs driven by the adoption of electronic trading (Pagano and Roell (1996), Tse and Zabotina (2001), Pirrong (1996), and Menkveld (2016)). In the same figure, we also show the actual spread value computed with intraday bid-ask effective spreads in corn futures. The estimated and actual series are remarkably different: HL estimates exceed effective spreads almost tenfold. Besides the difference in levels, the daily correlation between estimated and true spreads is not statistically different from zero.



Notes: The TAQ data comes from the CME Group BBO dataset. Prices used are always with respect to the nearby corn futures contract, with rollover on the first trading day of the expiration month. The effective spread is computed as $e_i \equiv 2|P_i - M_i|/M_i$, where P_i and M_i are the trade price and the midpoint (arithmetic average) of the outstanding bid-ask spread, respectively, at second i. A daily effective spread is a trade-weighted average of e over all seconds in a given trading day. Daily high-low (HL) spread estimates are computed according to the formula given in the main text with end-of-day data from CME. Daily negative estimates are set to zero and we plot monthly averages. The implementation of HL also uses the liquidity and overnight return *ad hoc* adjustments.

FIGURE II
The Failure of Simple BAS Estimators: An Example With Corn Futures

What could explain such discrepancy between estimated and actual spreads in corn futures, given that simple spread proxies on average work well in US stock data? A possible explanation is that model assumptions fail to hold in the data. One of the many virtues of simple bid-ask spread estimators lies in the transparency of the simplifying and identifying assumptions they rely on. Two violations that are easily verifiable are the presence of infrequent trading and overnight returns. Therefore, if corn futures are thin markets and there are large overnight returns, BAS proxies may perform poorly. These are very unlikely hypotheses not supported in the data. First, corn contracts are one of the most heavily traded commodity futures, so market thickness is not an issue. Nearby corn futures traded at a daily volume exceeding 150 thousand contracts from 2010 to 2018, which is about the same volume as gold futures. Second, even under large overnight returns, which are 0.57% versus 1.10% close-to-close returns in the sample, HL has *ad hoc* adjustments that should correct for model violations, both with respect to infrequent trading and large non-trading period price movements. Third, the CHL measure is fairly robust to infrequent trading and it is not affected by overnight returns.

Alternatively, an important difference between corn futures and the standard setting where HL and CHL are usually evaluated is the average spread size. The average effective spread in corn futures from 2008-2016 is less than 0.1%, which is essentially at the lower bound implied by the transaction tick. Although this effective spread aligns with spreads of most stocks, financial and commodity futures, the misbehavior reported in Figure (II) goes unnoticed in performance studies based on average measures because their usual setting contains a sizable fraction of illiquid assets. When the spread is large, the HL and CHL measures tend to perform better, thus driving high average cross-sectional correlations between the proxies and the effective spread.

3 Performance of simple bid-ask spread estimators

3.1 Base simulation results

Given that the corn price data is unlikely to violate the assumptions of the high-low and close-high-low estimators, alternatively, the spread size could be driving the large bias in both HL and CHL series. To investigate the accuracy of estimated spreads to the underlying true spread size, we set a standard simulation environment with financial price series data satisfying simple BAS estimator model assumptions, and then analyze the consequences for the estimators' behavior

in terms of spread size variation.

We generate 10,000 twenty-one-day months of efficient price data. Prices are created for 390 minutes m on each day, following $\mathcal{P}_m = \mathcal{P}_{m-1} + x\sigma_m$, where $x \sim \mathrm{N}(0,1)$, and $\sigma_m = \sigma/\sqrt{390}$ is the standard deviation per minute. The daily standard deviation of efficient price returns is set constant to $\sigma = 3\%$. Observed prices are then calculated after compounding or discounting true prices by half the spread S, as in Equation (1). For every disjoint interval of 390 minutes, we compute daily high, low and closing observed prices. At the beginning of each month, we normalize the true price as $\$(\ln 100)$. This data generating process is similar to the simulation environments in Corwin and Schultz (2012) and Abdi and Ranaldo (2017). The main difference is the addition of spread levels below 0.5% and the frequency of negative point-estimates.

The top portion of each subpanel of Table (I), when *ex ante* spread values are lower than 1%, is of particular interest. Subpanel A corresponds to near ideal conditions for implementation. Prices are observed "continuously", there are no overnight returns, and the true spread is constant over time. Two major regularities emerge. First, the bias in the average spread estimate increases as the level of the true spread being proxied falls below 1%. From 1% to 0.5%, the high-low measure is slightly more biased upwards than the CHL estimator. The pattern reverses as the spread shrinks, and the bias of CHL average estimates for a spread level of 0.1% is 12-fold higher than the true value. The other clear result displayed in the table aligns with the direction of bias. The frequency of point-estimates that turn out to be negative also grows as the spread size decreases, and accounts for nearly half of CHL estimates in the 210,000 trading days simulated. The dispersion of daily estimated spreads also follows the tendency of bias, with the same reversal of performance from both estimators at the spread size of 0.5%.

Subpanel B introduces infrequent trading as a model violation while maintaining the other ideal settings. We include rarely observed trades to emphasize that qualitative results are identical for the CHL measure (and very similar for the HL proxy). Decreasing the probability of observing a particular trade would lead to similar conclusions. The simulation results are important to suggest a connection between the frequency of negative spread estimates, bias in the average estimate, and true spread size. We explore these connections in following sections.

TABLE I

Simulation Results:

HL and CHL Estimates As the Spread Size Changes

	Bias		Negative		RMSEs	
	HL	CHL	HL	CHL	HL	CHL
Subpanel A. Near-ideal conditions						
0.10% spread	1.12%	1.22%	42.35%	49.37%	1.16%	1.29%
0.25% spread	1.05%	1.08%	40.87%	49.36%	1.10%	1.15%
0.50% spread	0.94%	0.85%	38.41%	49.03%	1.00%	0.95%
1.00% spread	0.75%	0.46%	33.84%	47.58%	0.84%	0.62%
3.00% spread	0.19%	−0.49%	18.47%	32.34%	0.52%	0.74%
5.00% spread	−0.11%	−0.60%	8.94%	13.19%	0.57%	0.88%
8.00% spread	−0.32%	−0.38%	3.81%	2.77%	0.69%	0.77%
Subpanel B. Each trade is visible with a chance of 10%						
0.10% spread	0.84%	1.22%	48.54%	49.60%	0.85%	1.29%
0.25% spread	0.76%	1.07%	47.02%	49.54%	0.78%	1.15%
0.50% spread	0.61%	0.85%	44.54%	49.08%	0.67%	0.49%
1.00% spread	0.38%	0.45%	39.62%	47.54%	0.50%	0.63%
3.00% spread	−0.29%	−0.47%	23.21%	32.30%	0.54%	0.73%
5.00% spread	−0.68%	−0.60%	12.19%	13.90%	0.87%	0.88%
8.00% spread	−0.87%	−0.40%	5.05%	3.18%	1.14%	0.79%

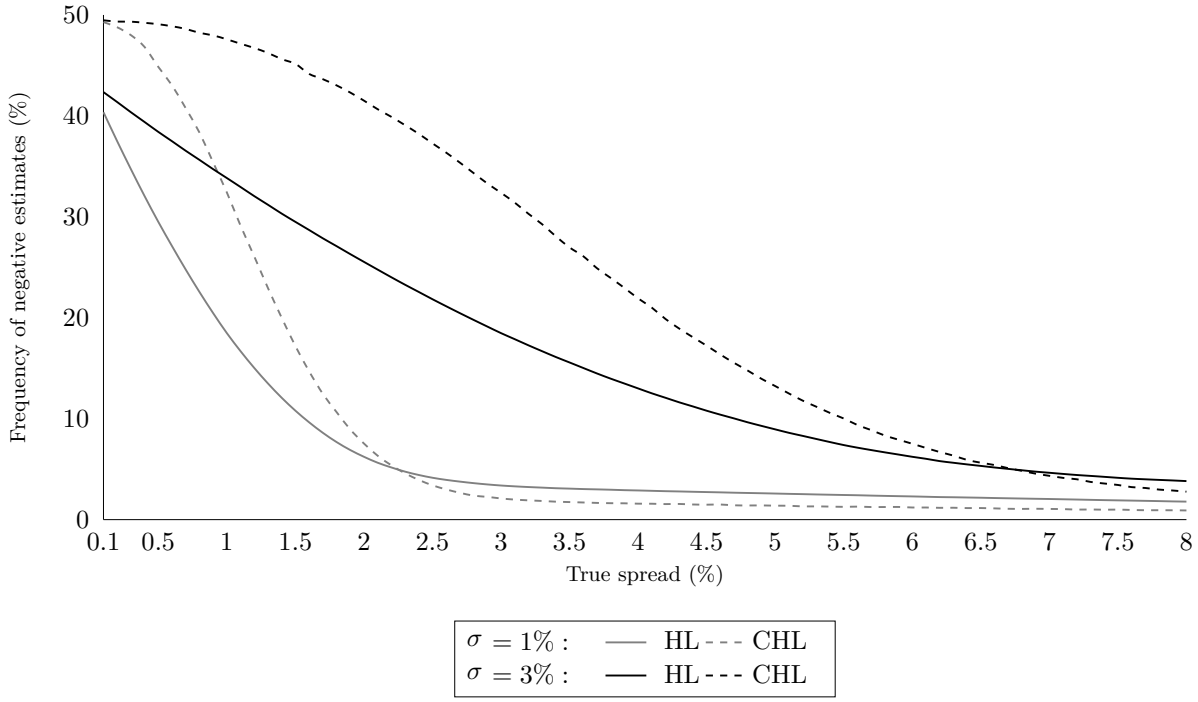
Notes: Values displayed for the bias are computed as the true spread minus the average spread estimate $\overline{\text{HL}}$ for the high-low estimator and $\overline{\text{CHL}}$ for the close-high-low estimator. Average estimates are calculated after imputing zeros when point-estimates are negative. There are 10,000 months of simulated efficient price data, each month with 21 days. Within a day, prices are observed for 390 minutes (the total discrete trading time) in Subpanel A and on average 39 times per day in Subpanel B where each trade has a probability of being observed set to 10%. Daily high, low and closing prices are derived using the selected true spread level. More details on the data generating process (DGP) can be found in the main text. The proportion of computed negative spread estimates is simply the frequency of $2\tanh(\alpha/2) < 0$ in the case of HL and $(c_t - \eta_t)(c_t - \eta_{t+1}) < 0$ for CHL.

3.2 Using negative estimates to infer true spreads

In spite of the usefulness of simulation exercises, the obvious limitation to drawing general conclusions is the inability to measure the sensitivity of the model to perturbations in unobservable variables. One of the most attractive characteristics from the family of simple bid-ask spread estimators lies in the transparency of its assumptions. Such feature allows us to easily disentangle the findings in Table (I) from the failure of model assumptions. The downside is missing one easily obtainable source of explanation of misbehavior (i.e. dropping a model assumption). We now turn our attention to the the joint-movement of the proportion of negative estimates and spread size. Since the fraction of negative estimated spreads is a mechanical consequence of point-estimate behavior, it may convey useful information on the spread.

In Figure (III), we take a closer look at the proportion of negative simulated estimates as the true spread varies, given the level of daily volatility. The values presented in Table (I) correspond to the appropriate plotted points for HL and CHL when $\sigma = 3\%$. Clearly the close-high-low measure is negative more often than its predecessor HL, until both frequencies converge to near zero for high spread values. The correlation between each proportion of negative spreads and true spread size is above 90%. We also plot the frequency of negative point-estimates measured with trading data generated under a daily volatility regime of $\sigma = 1\%$, which is fairly low for current financial markets. In any case, one could be interested in the effect of greater daily variation in the number of negative estimates. Under low daily volatility, the frequencies decay much faster, and spread sizes above 2% are sufficient to generate virtually only positive estimates. In this case, the correlation between spread size and negative spreads in each estimated series is about 70%. Notwithstanding, the strong relationship between spread size and the frequency of negative point-estimates, given the volatility level, remains.

We focus on the frequency of negative spreads not only because reducing unfeasible estimates is one of the main incentives for developing new spread proxies, but also because it is empirically observable. Prior to imputing zeros, the researcher can easily record the frequency of negative estimates. Thus, the frequency of negative spreads can be useful if we are able to provide a proper framework connecting spread size and volatility to the observance of negative estimates. In the next subsection, we address the question: Why are estimated spreads more often negative when



Notes: This figure compares the frequency of negative high-low (HL) and close-high-low (CHL) estimates from simulations assuming different values of true spread and volatility level. For each combination of spread and daily volatility, 210,000 days of trading data are generated according to the DGP described in the main text. The proportion of computed negative spread estimates is simply the frequency of $2\tanh(\alpha/2) < 0$ in the case of HL and $(c_t - \eta_t)(c_t - \eta_{t+1}) < 0$ for CHL.

FIGURE III
Frequency of Negative HL and CHL Spread Estimates
As the Spread Size Changes

the true spread is small?

3.3 Negativity of the high-low estimator

The HL estimator depends upon both the first and second-day ranges, r_t and r_{t+1}, and the two-day range r_t^*. Small perturbations at price boundaries have small effects on daily range values, but may alter the domain of the parameter γ and lead to considerable swings in the HL estimate. If changes in the second-day high, for example, are sufficiently large so that h_2 becomes higher than the previous day high, the mapping of γ shifts from h_1 to h_2, and the relationship between β and γ is updated. From the expression for α in (6), it is clear that the high-low estimator is positively defined only when $\beta > \gamma$.

Negative spread estimates do not arise only if the volatility-proportionality assumption embedded in (5) is violated. Consider the following example. Suppose the first-day range is $r_1 = 0.01$,

which is obtained when $(h_1, l_1) = (350, 345)$, and the second-day range is $r_2 = 0.27$, for $(h_2, l_2) = (401, 305)$. Clearly, the second day is much more volatile than the previous day, but the HL estimate is $0.09\% > 0$ over the two days. The simple explanation is that the square of the two-day range cannot be larger than itself added to the square of the first day range: $\gamma = r_2^2 < \beta = r_1^2 + r_2^2$. This happens because r_1 is enclosed in the second-day range. Note that it follows from this reasoning that the HL estimator can only be negative when the parameter γ combines boundaries from both consecutive days (i.e. either $h_1 - l_2$ or $h_2 - l_1$).

We use this fact in helping to derive a general condition which establishes the negativity of the high-low estimator.

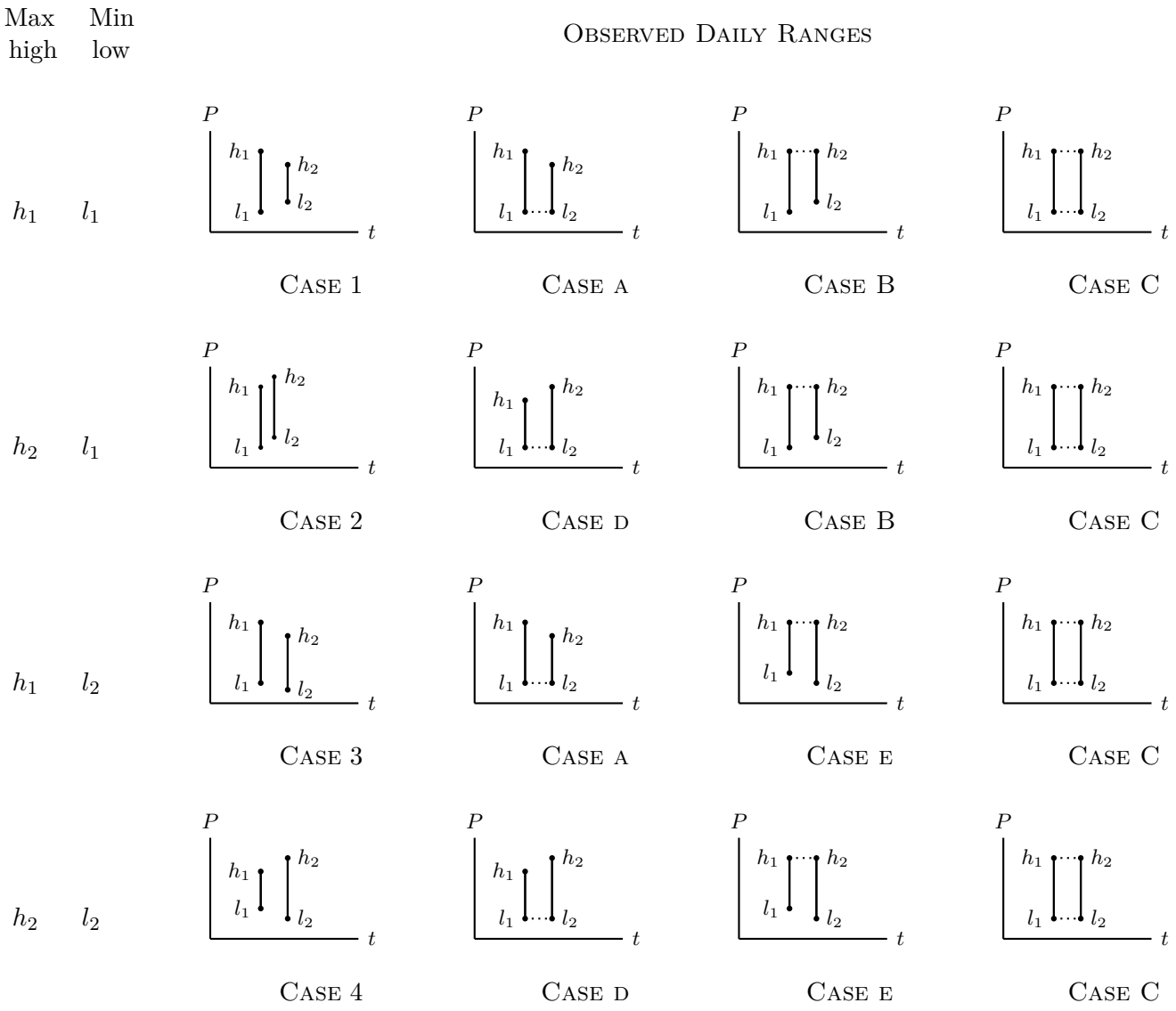
LEMMA 1. *The high-low spread estimator* (HL_t) *is negative if and only if the following negativity condition holds:*

$$|\eta_{t+1} - \eta_t| > \sqrt{r_t^2 + r_{t+1}^2} - \left(\frac{r_t + r_{t+1}}{2} \right)$$

where η_t is the (log) mid-range and r_t is the log range on day t.

In Lemma 1, we explore the relationship between the independence of the observed mid-range from the bid-ask spread bounce, relative to the "contamination" of the range by microstructure noise. In order to show that the inequality is necessary and sufficient to determine the negativity of the HL estimator, we first argue that the high-low spread estimator may only be negative in two cases, conditioned on the values taken by γ. These possibilities correspond to cases 2 and 3 in Figure (IV). In contrast, cases 1 and 4 in the figure illustrate when one daily range is enclosed in another day's range, and thus the HL is always positive. We then combine the two separate solutions into one, which yields the lemma. The proof is in the Appendix (A), and without loss of generality, it assumes there are no price ties, i.e. high and low daily prices over each two-day interval are not equal. In Figure (IV) we include the cases with price ties (A through E) to give a full description of all range combinations that generate each possible value of γ.[7] We also assume that the range is well defined: $h_t > l_t$ for all t trading days.

[7] In the Appendix (B), we show that the HL estimator is always positive when there are price ties.



Notes: The figure shows the possible relative positions of two-consecutive daily log ranges that generate the domain of γ. For $t = 1$, the parameter γ is defined as $\gamma \equiv r_1^{*2}$, where $r_t^* \equiv \max\{h_1, h_2\} - \min\{l_1, l_2\}$. For each of the four observable values of γ, there are four possible relative positions of daily ranges. One without price ties and three with at least one price tie. For example, when the maximum high is h_2 and the minimum low is l_1 (second row), a first possibility is that $l_1 < l_2$ and $h_2 > h_1$ (case 2), which is without price ties, in addition to $l_1 = l_2$ and $h_1 < h_2$; $l_1 < l_2$ and $h_1 = h_2$; $l_1 = l_2$ and $h_1 = h_2$.

FIGURE IV

Combinations of daily high and low prices

Lemma 1 provides the framework for two important mechanisms that affect the empirical use of HL. The first is the difference of noise in the daily mid-range relatively to the range. To simplify, rewrite the negativity condition as $f(\eta_t^o, \eta_{t+1}^o) > g(r_t^o, r_{t+1}^o)$. Because mid-range prices are independent of the bid-ask bounce, but changes in *ex-ante* spreads "contaminate" the range, variations in the spread level only affect g. Intuitively, g should increase in the spread level S (given $g > 0$), so that low transaction costs would lead to a wide range of two-day prices resulting in negative spread estimates.[8] On the other hand, both mid-range and range capture daily volatility, so that f and g vary with changes in σ^2. This is the second mechanism.

Consider the example on the left-hand side of Figure (V), which corresponds to case2. With a partial overlap between r_t and r_{t+1}, the high-low estimator is not necessarily negative. Particularly, it is straightforward to show that some combination of high and low prices will produce a positive HL estimate when $l_{t+1} \in (l_t, h_t)$ and $h_{t+1} > h_t$. When we decrease the spread S by half (right-hand side panel), the new relative position of the second-day low, $l_{t+1} > h_t$, implies that the HL estimate is negative.

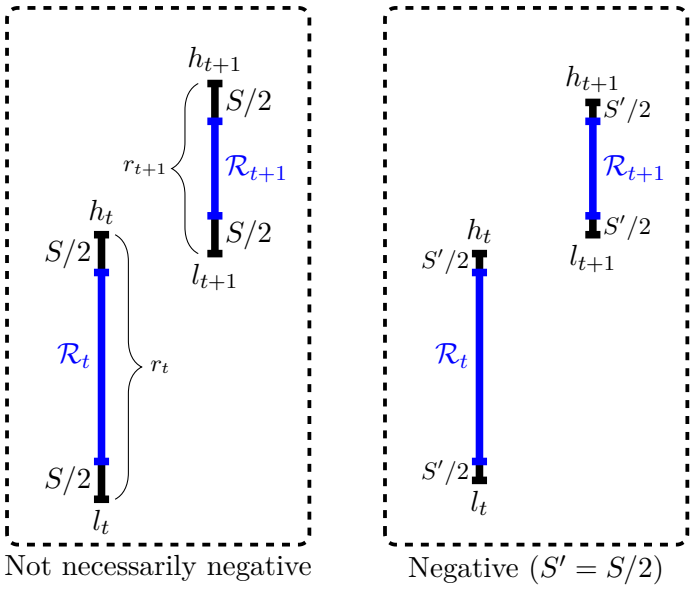


FIGURE V
NEGATIVITY INDUCED BY THE SPREAD SIZE IN THE HL ESTIMATOR

In Figure (III), the frequency of negative estimates is extremely high around regions of small

[8] It is easy to check that $\sqrt{r_t^2 + r_{t+1}^2} > \frac{r_t + r_{t+1}}{2}$.

spreads, even for different choice values of daily volatility. As the spread increases, volatility plays a bigger role in determining negative estimates, up to the point where the frequency of negative estimates becomes invariant to volatility, which happens when the spread level is considerably large. These two patterns combined suggest that the effects of σ^2 in the negativity condition are conditioned on the spread level. We formalize the observable effects of spread and volatility levels in the negativity condition of the HL measure in the proposition below.

PROPOSITION 1. *Let $f \equiv f(\eta_t, \eta_{t+1})$ and $g \equiv g(r_t, r_{t+1})$ be functions that define the left-hand and right-hand sides, respectively, of the negativity condition for the high-low spread estimator* (HL_t)*:*

$$f(\eta_t, \eta_{t+1}) > g(r_t, r_{t+1}) \tag{9}$$

For ex-ante changes in the bid-ask spread level S and variance σ^2, the following relationships hold:

$$\frac{\partial f}{\partial S} = 0, \quad \frac{\partial g}{\partial S} > 0, \quad \frac{\partial E[f]}{\partial \sigma^2} = (2 - 2\ln 2), \quad \frac{\partial E[g]}{\partial \sigma^2} > 0, \quad \text{and} \quad \frac{\partial^2 E[g]}{\partial \sigma^2 \partial S} > 0 \tag{10}$$

We let two arbitrary functions f and g represent the inequality condition for the HL estimator to make the proposition more intuitive. The left-hand side f contains only mid-range prices, η, while the right-hand side g contains only ranges r. The greater f is compared to g, the easier the negativity condition is attained.

PROOF. Because the efficient mid-range is identical to the observed mid-range, the left-hand side of the inequality does not depend on the spread level. That is, volatility estimated with squared returns of mid-prices is independent of the spread level. Thus $\partial f / \partial S = 0$ follows. On the other hand, the right-hand side of the inequality is a function of the observed range, $g(r_t, r_{t+1})$, which depends on the spread. If we rewrite observed ranges in terms of efficient ranges, the modified function g depends on S in the following way:

$$g \equiv \sqrt{(\mathcal{R}_t + S)^2 + (\mathcal{R}_{t+1} + S)^2} - \left(\frac{\mathcal{R}_t + \mathcal{R}_{t+1} + 2S}{2} \right)$$

and therefore

$$\frac{\partial g}{\partial S} = \frac{(\mathcal{R}_t + S) + (\mathcal{R}_{t+1} + S)}{\sqrt{(\mathcal{R}_t + S)^2 + (\mathcal{R}_{t+1} + S)^2}} - 1 > 0$$

for well-defined efficient ranges. That implies that the right-hand side in the negativity condition increases in the spread; hence as the spread widens, $\sqrt{\beta}$ relatively exceeds $(r_t + r_{t+1})/2$, so that the inequality becomes more difficult to be attained. For relatively greater spreads, the HL estimator is more likely to be positive.

The second part of Proposition 1 is also simple. Denote $r_{t+1} = \kappa r_t$, $\kappa > 0$, but not necessarily bounded by 1. After squaring both sides of the negativity condition with the proper substitutions for r_{t+1}, we have

$$(\eta_{t+1} - \eta_t)^2 > \frac{1}{4}\left(\kappa - 2\sqrt{\kappa^2 + 1} + 1\right)^2 r_t^2$$

for which we can replace the observed range with the efficient range. To simplify the calculations, assume $\overline{\kappa} = 1$. Under the maintained hypothesis of constant volatility, daily observed range values will be very close, and κ will differ from the unity when the spread is very small. The average value of κ in 210,000 days of simulated data when $\sigma = 3\%$ and with an infinitesimal spread is 1.09. Table (A.1) in the Appendix supports our simplification as fairly representative of historical stock data. With $\overline{\kappa} = 1$, the term multiplying r_t^2 collapses to $(3 - 2\sqrt{2})$ and after taking expectations of both sides, we arrive at

$$E\left[(\eta_{t+1} - \eta_t)^2\right] > (3 - 2\sqrt{2})E\left[\mathcal{R}_t^2\right] + 2S(3 - 2\sqrt{2})E\left[\mathcal{R}_t\right] + (3 - 2\sqrt{2})S^2.$$

Direct substitutions for the moments above using (3) and (8) yield

$$\left(2 - \frac{k_2}{2}\right)\sigma^2 > (3 - 2\sqrt{2})k_2\sigma^2 + 2(3 - 2\sqrt{2})k_1 S\sqrt{\sigma^2} + (3 - 2\sqrt{2})S^2. \tag{11}$$

Since the left-hand side of (11) is now $E[f]$, we have $\partial E[f]/\partial\sigma^2 > 0$. Similarly, the right-hand side is given by $E[g]$; hence $\partial E[g]/\partial\sigma^2 > 0$ follows. The growth rate of $E[g]$ in σ^2 may exceed $2 - k_2/2$, which largely depends on the magnitude of S. Since $\partial^2 E[g]/\partial\sigma^2\partial S > 0$, smaller spreads will reduce $E[g]$ even while volatility increases, so that the net effect of volatility depends on the relative size between S and σ^2. This completes the proof of the proposition.

There are three main implications to Proposition 1. First, smaller unobserved spreads induce more negative high-low estimates. Second, increasing volatility has an ambiguous effect on the proportion of negative estimates, as $E[g]$ may or may not exceed $E[f]$. Third and relatedly, S regulates the relative contribution of σ^2 in $E[g]$. This implies that ultimately the spread is more important than volatility to determine when the high-low estimator is negative.

Reductions in the *ex ante* spread increase the frequency of negative spreads through g and $E[g]$. Because S acts as a limiting factor to the growth of $E[g]$ with respect to σ^2, we should expect a high proportion of negative spreads for a narrow spread, regardless of the volatility level. In the limit, as $S \to 0$, the right-hand side of (11) converges to $(3 - 2\sqrt{2})k_2\sigma^2$, which is the minimum of $E[g]$ with respect to S and smaller than $2 - k_2/2$. In the theoretical case, the expected frequency of negative spreads is 100%. In practice, either with simulated or actual data that can only be observed at a discrete set of times, the proportion of negative estimates is maximized when the underlying asset is very liquid. In Figure (III), when the true spread is 0.1%, increasing volatility from 1% to 3% produces almost no impact in the proportion of negative HL estimates.[9]

Observable differences in the frequency of negative estimates resulting from varying σ^2 start to appear as the spread grows. Locally widening the spread enables volatility to contribute toward the growth rate of $E[g]$. The difference $E[f] - E[g]$ increases in σ^2 up to a spread level where $E[g]$ is greater than $E[f]$. At this point, the frequency of negative estimates is minimized, and different levels of volatility are again redundant to (11).

3.4 Negativity of the close-high-low estimator

We follow the same steps for the close-high-low estimator. Even though the CHL measure does not depend on the range, it does depend on the mid-range. Moreover, our simulation results show a behavior dynamic of CHL similar to HL with respect to negative spreads, true spread size, and bias. This suggests a common channel through which the spread-to-volatility ratio affects both proxies. From the expression for CHL_t, the straightforward negativity condition of the close-high-low estimator is:

$$c_t \in (\eta_t, \eta_{t+1}) \tag{12}$$

[9] The scaling factors k_1 and k_2 represent moments of the range of a standard Brownian motion and therefore are not numerically precise (Christensen and Podolskij (2007)) to evaluate the range observed discretely.

when the second-day mid-range is above the first-day mid-range, and $c_t \in (\eta_{t+1}, \eta_t)$ in the symmetric case. When the price variation across two days is large enough (measured by the difference in mid-prices), a wide range of values for c_t yields negative CHL spread estimates. We can rewrite (12) as (by analogy the symmetric case follows)

$$\eta_{t+1} - \eta_t > c_t - \eta_t \tag{13}$$

and similarly to HL, define an inequality condition to investigate the effects of spread and volatility levels in generating negative spread estimates.

PROPOSITION 2. *Let $f \equiv f(\eta_t, \eta_{t+1})$ and $v \equiv v(c_t, \eta_t)$ be functions that define the left-hand and right-hand sides, respectively, of the negativity condition for the close-high-low spread estimator* (CHL_t)*:*

$$f(\eta_t, \eta_{t+1}) > v(c_t, \eta_t) \tag{14}$$

For ex-ante changes in the bid-ask spread level S and variance σ^2, the following relationships hold:

$$\frac{\partial f}{\partial S} = 0, \quad \frac{\partial E[v]}{\partial S} > 0, \quad \frac{\partial E[f]}{\partial \sigma^2} = k_3, \quad \frac{\partial E[v]}{\partial \sigma^2} = \frac{k_3}{2} \quad \text{and} \quad \frac{\partial^2 E[v]}{\partial \sigma^2 \partial S} = 0. \tag{15}$$

Proposition 2 is intended to mimic the structure used in Proposition 1. The generic function f is intentionally identical to f in the HL estimator. The proof, however, is slightly different from Proposition 1. Since we do not assume the value of the trade indicator q_t in $c_t = \mathcal{C}_t + q_t S/2$, we can only work with the expected value of v, without assuming a deterministic q_t. We also use moments derived for $E\left[\mathcal{C}^a \mathcal{H}^b \mathcal{L}^c\right]$ in Garman and Klass (1980). The proof is in the Appendix (A).

The interpretation of Proposition 2 is very similar to Proposition 1, although simpler. Decreasing the spread size lowers the expected value of the right-hand side of the negativity condition, contributing to more negative estimates. When the spread is small, increases in σ^2 cannot induce $E[f] < E[v]$, which explains why in Figure (III) the proportion of negative CHL estimates is around 50% when $S = 0.1\%$, regardless of volatility size. Moreover, for small spreads, any value of σ^2 increases $E[f]$ faster than $E[v]$ – a pattern that is eventually reversed when the spread is large enough so that the right-hand side becomes greater than $E[f]$. For completeness, we also provide an

alternative result (Proposition 3 in the Appendix) that shows why smaller spreads induce negative CHL estimates more often.

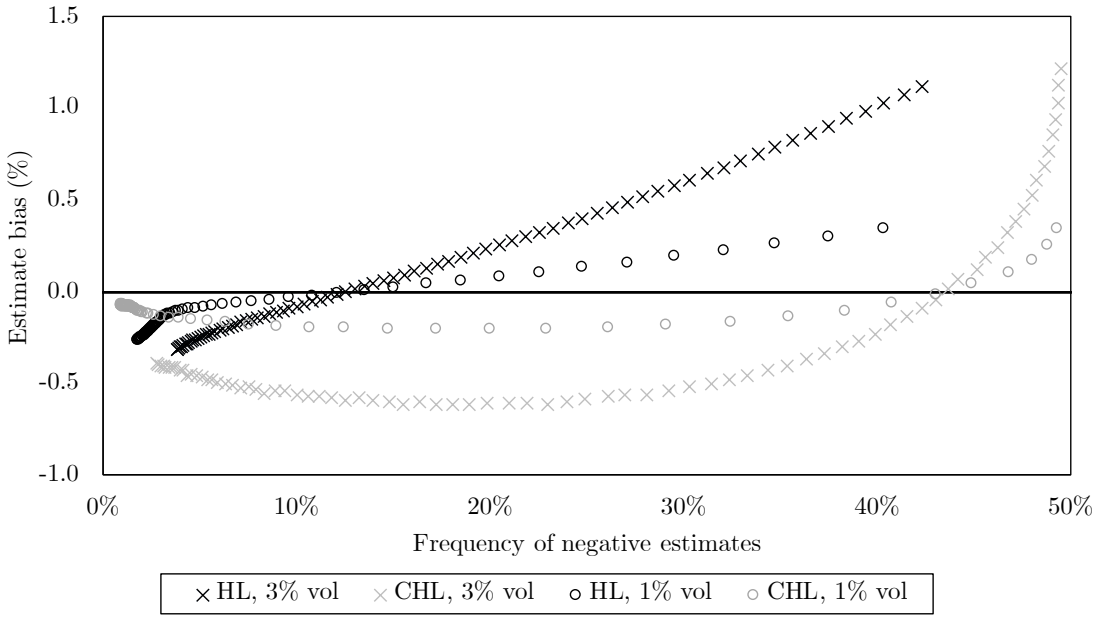
4 Sources of bias in simple bid-ask spread estimators

It is important to understand the mechanisms that induce negative estimates of high-low and close-high-low spreads for two complementary reasons. First, negative estimates are undesirable from an empirical perspective. Some adjustment procedure must be chosen, and ultimately more negative spreads reduce the sample size of "well-behaved" estimates. Common adjustment procedures to deal with negative spreads usually include zero imputation, missing-variable-type treatment, or simply averaging over both negative and positive spreads. The second justification follows our preliminary evidence in Section (3) that how often negative estimates occur and how much bias average BAS estimates suffer from are intertwined. Since the frequency of negative spreads can be observed in practice, it may convey useful information regarding the lack of accuracy one incurs when implementing simple bid-ask spread estimators.

The average estimates computed in the simulated environment included the *ad hoc* adjustment of imputing zeros prior to obtaining monthly averages and then the average estimate itself. The average spread is therefore an average of non-negative estimated spreads and imputed zeros, both for $\overline{\text{HL}}$ and $\overline{\text{CHL}}$. Figure (VI) provides greater detail on the relationship between average spreads and the frequency of negative estimates. There is a clear positive association between both in the case of HL. For the CHL estimator, the relationship displays convexity, with a positive association when the proportion of negative spreads is at least 40%.

The impression that "some degree of negative estimates is good", in a sense that the bias is zero when the proportion of negative spreads is positive, is just mechanical. If point-estimates are accurate, imputing zeros should bias the average estimates toward zero (as in Figure (I)). In that case, we should expect the opposite pattern from the one displayed in the figure: the more often zeros are imputed, the smaller the average spread becomes, and as a consequence, the overall bias is negative. Considering that the bias in the average spread remains positive after imputing zeros, it must be that positive point-estimates suffer from

upward bias. Further, if the average estimator bias continues to increase even with relatively more zeros being imputed, the bias in positive spreads must also continue to increase.



Notes: The figure plots the frequency of negative estimated spreads and the bias in the average spread estimate for the high-low (HL) estimator and close-high-low (CHL) estimator. Each observation represents the average spread computed with point-estimates calculated from 210,000 simulated days of transaction data. For each point, the chosen level of true spread differs, which generates a different level of negative estimates and bias. For example, the first cross for HL with $\sigma = 3\%$ indicates the bias and negative frequency when $S = 8\%$. The last cross in the same series is obtained when $S = 0.1\%$. Along the horizontal axis, the true spread level decreases from 8% to 0.1% with decimal variations. The proportion of negative spread estimates is the frequency of $2\tanh(\alpha/2) < 0$ in the case of HL and $(c_t - \eta_t)(c_t - \eta_{t+1}) < 0$ for CHL. Bias is measured as $\overline{\text{HL}} - S$ and $\overline{\text{CHL}} - S$, where average estimates have negative point-estimates set to zero.

FIGURE VI
Relationship Between Bias and Proportion of Negative Estimates

In this section we show that the HL and CHL average spread estimators suffer from upward bias mainly introduced from positive point-estimates: the "well-behaved" high-low and close-high-low estimates that are non-negative. The sample average of positive estimates overestimates the true spread, and the degree of overestimation increases as the true spread size decreases. We show this must be the case since reducing the number of generated positive spreads, thereby increasing the number of negative estimates, does not drive the bias alone. The central insight from the analysis is that, *ceteris paribus*, what induces more negative estimates is a true spread level that progressively becomes smaller. Therefore, if observing a high frequency of negative point-estimates identifies a small true spread, and a small true

spread induces upward bias in the average of positive estimated spreads, the proportion of negatives can be used to infer the magnitude of biases.

4.1 A problem of zeros

We illustrate the sources of bias affecting simple BAS proxies with the average high-low estimator, $\overline{\text{HL}}$, with $J = T$. The average estimator consists of both non-negative and negative point-estimates:

$$\overline{\text{HL}} \approx \frac{1}{2}\left(\underbrace{\frac{\sum_{k=1}^{T-N} \text{HL}_k^*}{T-N}}_{\text{Positive}} + \underbrace{\frac{\sum_{n=1}^{N} \text{HL}_n^o}{N}}_{\text{Negative}} \right) \tag{16}$$

where $(T-N)^{-1} \sum_{k=1}^{T-N} \text{HL}_k^*$ is the average of non-negative point estimates, whom we call positive term and similarly $N^{-1} \sum_{n=1}^{N} \text{HL}_k^o$ is the average of negative estimates, thereby the negative term. Assume the positive term is an unbiased estimate of the true spread S regardless of its sample size $T-N$ and the spread level. The average of any number of estimated non-positive spreads that the researcher observes equals the latent spread, for any spread size. Also assume that positive spreads are symmetric around S and have a small degree of dispersion. Under this setting, the best way to deal with the negative term is to discard it and average $\overline{\text{HL}}$ only over HL^*. Imputing zeros biases the monthly estimate toward zero, and if N is large enough, computing the negative estimates in the average estimator may cause $\overline{\text{HL}}$ to become negative.

Consider now that the positive term in (16) converges to S only when $T-N$ is large enough. Of course, $T-N$ is limited by the particular context. As a consequence, a greater frequency of negative estimates *per se* is associated with bias in the positive term since it necessarily decreases the size of $T-N$. The magnitude and sign of bias are indeterminate, but under our assumptions, are likely to be small. Independent of the bias direction, imputing zeros in the negative term shifts the average estimator downwards, which may balance out $\overline{\text{HL}}$ if the bias in HL^* is positive, or worsen the downward bias if the contrary. Similar to before, computing the negative spreads without any adjustment produces a somewhat similar effect, if the distribution of negative spreads is negatively skewed and not very disperse –

i.e. there are no excessively negative point-estimates.

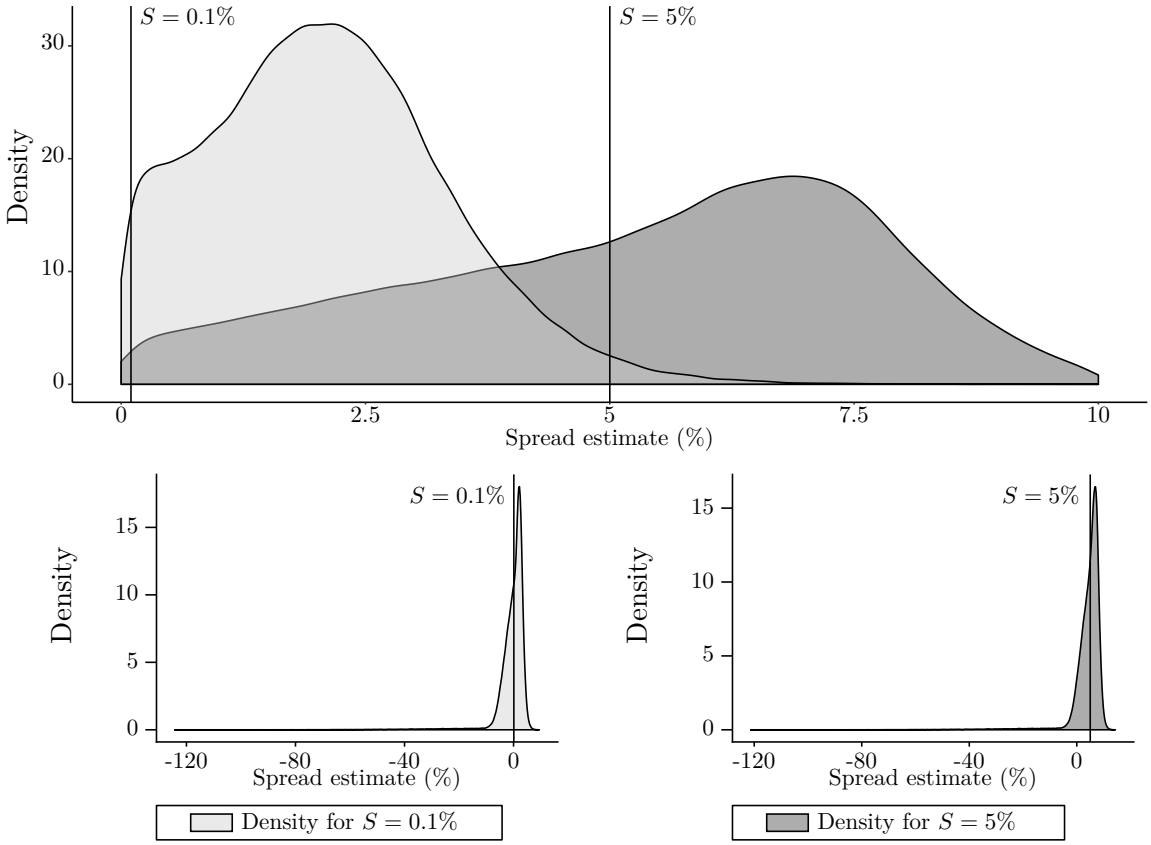
In Section (3) we demonstrated what changes the dimension of N. When the true spread size is small, the number of negative spread occurrences is likely to increase. If a greater N induces bias in the positive term of (16), a small spread size indirectly induces bias in the average estimator through the frequency of negative spreads. This is the first channel through which the size of the unobserved true spread affects $\overline{\text{HL}}$. To make things more realistic, let the accuracy of the positive term HL^* also depend on the spread size. For a fixed $T-N$, the convergence $(T-N)^{-1}\sum_{k=1}^{T-N}\text{HL}_k^* \to S$ only takes place if S is large enough.[10] For small spreads, the positive term overestimates S. This constitutes the second channel through which the spread size could bias $\overline{\text{HL}}$. Although we do not offer a formal statement as provided for the first channel, we argue that the direct effect of S on the bias in HL and CHL positive estimates is sizable. That is because the first channel cannot account for most of the bias in HL^*.

In the top portion of Figure (VII), we show the density implied by HL point-estimates in the simulation from Section (3) when $\sigma = 3\%$. The distribution in light gray is obtained for a true spread size of 0.1% and the one in dark gray when $S = 5\%$. The densities correspond only to HL^* across all simulated months, so the total N in the light gray area is smaller than the number of observations that generated the dark gray density. The sample average of positive-only spreads when $S = 0.1\%$ is 2.11%. Even though $T - N = 121,061$, random draws with different sizes from the subsample of positive spreads have very similar means. For example, a 1% sample draw ($T - N = 1,211$) yields an estimate of 2.10%, while a 10% draw returns an estimate of 2.12%. When the true spread equals 5%, $(T-N)^{-1}\sum_{k=1}^{T-N}\text{HL}_k^* = 5.37\%$, for $T - N = 192,217$. Again, reducing the sample size does not produce significant bias in the restricted HL^*: average estimates are 5.34% from a 1% random sample and 5.35% from a 10% subsample. The possibility that decreasing the sample size of positive estimates would alone explain the large positive bias in the positive term in (16) is not a consistent explanation.[11] Because a greater number of negative estimates identifies a small true spread

[10] This is arbitrary (the requirement on S could be the opposite to enable the convergence of the positive subsample to S); however, as we explain in sequence, this is the correct effect of the spread size on the positive term bias.

[11] One could be concerned whether the difference in bias between positive-only estimates with $S = 0.1\%$ and $S = 5\%$ is not explained by the differences in subsample sizes, 121,061 and 192,217, respectively. The fact that increasing $T - N$ from roughly a thousand to 121,061 does not alter $(T-N)^{-1}\sum_{k=1}^{T-N}\text{HL}_k^*$ suggests that if it would be possible to observe a decrease in N not caused by augmenting S, so that $T - N$ becomes 192,217 even when $S = 0.1\%$, the sample mean of positive spreads HL^* would not suddenly converge to S.

(given that efficient ranges and daily standard deviation of efficient prices remain the same), the bias in the positive term must be attributed to the change in the spread size.



Notes: The top part of the figure displays the empirical distribution of positive estimated high-low spreads under two different assumed spread sizes: 0.1% and 5%. For compactness, the distribution is contained in the domain $[0, 10\%]$. The bottom part displays each full distribution separately. The one on the left include all HL point-estimates generated when the true spread is 0.1% and the one on the right when the true spread is 5%. The point-estimates are computed from the simulation described in Section (3).

FIGURE VII
Distribution of Simulated HL Estimates

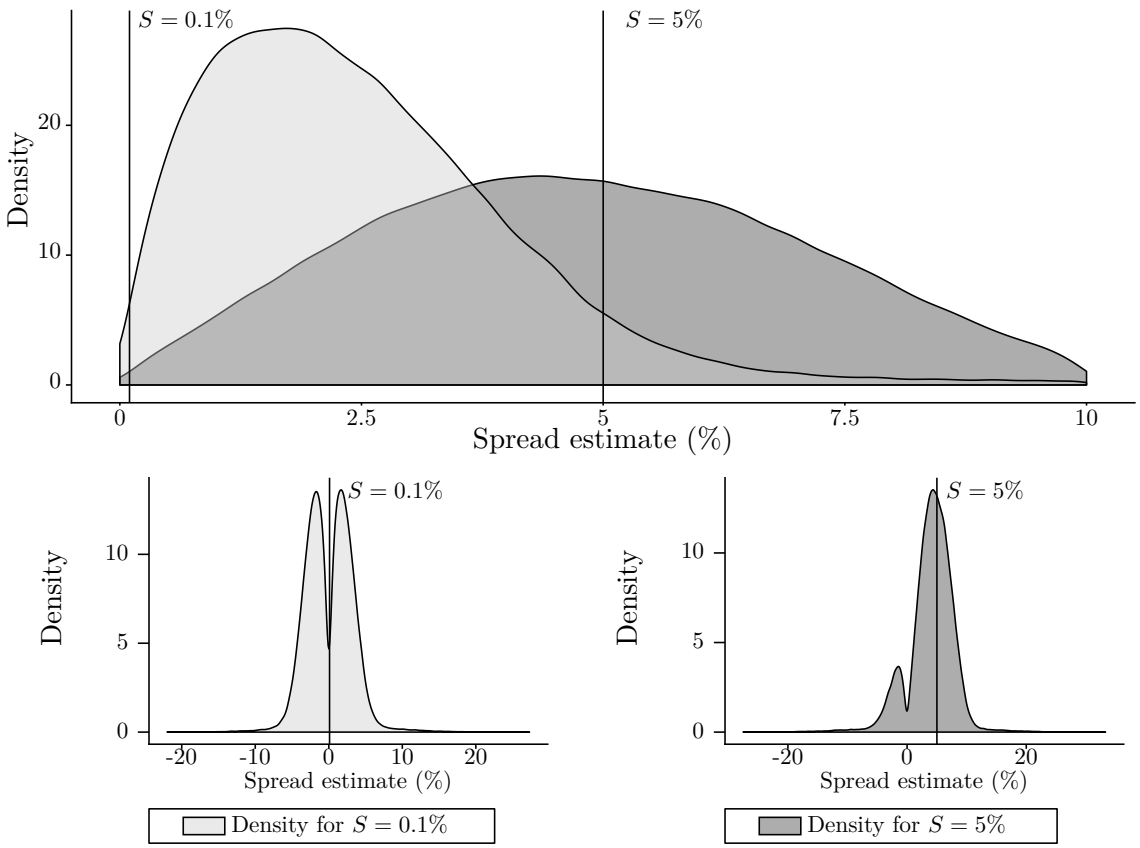
If the true spread is small − say 0.1% − the average of positive spread estimates will have a large upward bias. In addition, a small spread causes the frequency of negative estimates to increase. That reduces the sample size of positive daily estimated spreads, and may induce more upward bias in the average over positive estimates, although such bias is probably very small. Imputing zeros helps in reducing the overall bias particularly if the true spread size is small. The great number of negative estimates converted into zero weighs down the average HL estimator. This explains why Corwin and Schultz (2012) conclude that

the adjustment maximizes the performance of the HL estimator. As mentioned before, if positive-only spreads were accurate, imputing zeros would actually distort $\overline{\text{HL}}$. Nonetheless, as seen in Figure (VI), setting negative spreads to zero does not eliminate the average estimator bias. Could the use of negative estimates do so?

The bottom part of Figure (VII) displays the totality of point-estimates, both negative and non-negative spreads. The first pattern in the distributions generated when $S = 0.1\%$ and $S = 5\%$ is the long tail to the left that includes estimates of less than -100%. If negative spreads had negative skewness and a very short tail, they could help in alleviating the bias from the positive term (ignoring the question whether they make economic sense). Using negative spreads lowers the average $\overline{\text{HL}}$ by more than simply imputing zeros because there are so many large (in absolute value) estimates. For $S = 0.1\%$, a full sample average spread is -0.73%. The main difference between the two bottom graphs in Figure (VII) is the portion of positive spreads below each respective true spread line. Point-estimates lack precision regardless of the true spread level. As a consequence, when the true spread is very small, the effect of dispersion is relatively large because the assumed spread is close to zero.

The analysis is generally similar for the CHL estimator. However, there are two differences between the CHL estimator and HL shown in Figure (VIII) that are noteworthy. First, the subsample of positive estimates when $S = 5\%$ appears more symmetrical around S than the gray area in the previous figure. Secondly, as the spread size decays and the sample size of negative estimates rises, the distribution of non-positive spreads resembles a mirror image of the subsample with positive spreads. The full sample average of CHL for $S = 0.1\%$ is 0.08% − much more accurate than the 1.32% average with imputed zeros. In this particular case, including the negatives in the average estimator helps in reducing the bias. The drawback is that the procedure cannot be generalized: the positive-and-negative-spread mean for $S = 5\%$ is 4.00%, while the average with imputed zeros is 4.41% and with positive-only estimate is even more precise: 5.07%.

In this section we showed that both the HL and CHL proxies are affected by the underlying true spread size. When the latent spread is small − lower than 0.5% − the average over positive spread point-estimates suffers from large upward bias. Concomitantly, the proportion of negative, misbehaved estimates is also high, since at small true bid-ask



Notes: The top part of the figure displays the empirical distribution of positive estimated close-high-low spreads under two different assumed spread sizes: 0.1% and 5%. For compactness, the distribution is contained in the domain $[0, 10\%]$. The bottom part displays each full distribution separately. The one on the left include all HL point-estimates generated when the true spread is 0.1% and the one on the right when the true spread is 5%. The point-estimates are computed from the simulation described in Section (3).

FIGURE VIII
Distribution of Simulated CHL Estimates

bounce sizes it is easier for both HL and CHL to return negative spreads. The greater proportion of misbehaved estimates does not appear to induce bias in the positive term of (16) (or the equivalent version for CHL) by reducing the number of "sampled" positive estimates. Therefore the bias in average estimators must be explained by the direct effect of the magnitude of S. Accordingly, a high frequency of negative spreads identifies the bias in the average positive-only-spread estimator, because it reveals a small magnitude of the unobserved true spread. The final average estimates $\overline{\text{HL}}$ and $\overline{\text{CHL}}$ can be precise when S is relatively large, since dealing with negative estimates is irrelevant and sample averages of positive-only spreads are fairly accurate. For small S values, different choices of how to deal with negative estimates do not improve $\overline{\text{HL}}$; they may however increase the accuracy of

$\overline{\text{CHL}}$.

In the next section, we test whether the general predictions about the behavior of simple bid-ask spread estimators in a highly liquid market hold, and whether a predicted large bias is properly identified by a large proportion of negative estimates.

5 Empirical application

In this section we test the performance of the HL and CHL measures against high frequency data which allows us to observe actual trading costs. We choose corn futures prices since effective spreads in this market are fairly small and representative of reasonably liquid markets (Wang et al. (2014)). The data used is the Best Bid Offer (BBO) dataset from the electronic trading system CME Globex for corn futures and spans 2008-2016. We construct daily effective spreads following Section V in Corwin and Schultz (2012) as closely as possible.

In the BBO top-of-book database, the prevailing highest bid and lowest ask quotes are time-stamped to the nearest second and the corresponding trade receives a unique serial number. New, better pairs of bid and ask quotes, as well as changes in the number of outstanding contracts to be negotiated at current prices, introduce unique bid-ask pairs, along with bid, ask and trade sizes, and trade price observations. These new quotes receive an identification number that is incremental with respect to the previous sequence number, always preserving the ordering of most-recent quotes. Quoted prices vary one quarter of a cent per bushel, and prices are always from the nearby contract. Since futures contracts expire, we need to splice contracts with different maturities when the first nearby contract expires. We roll contracts on the first day of the expiration month.

It is common to observe multiple quotes for the same time-stamp in the data. We select the last trade within each second (the highest sequence number, given a same-time-stamp-interval), and construct the daily effective spread measure as: $e_i \equiv 2|P_i - M_i|/M_i$, where P_i is the trade price at second i, and M_i the midpoint (arithmetic average) of outstanding bid and ask spreads at the same second. The average effective spread is given by the trade-volume-weighted average of all trades within a day. The quoted BAS is simply the difference

between ask and bid quotes at every second. We average BAS values over the trading day to obtain daily BAS in cents per bushel. Lastly, intraday volatility is measured by the standard deviation of M_i across all seconds in a given day.

Table (II) presents the main results comparing simple bid-ask spread estimators to actual trading costs data. We report summary statistics for different versions of *ad hoc* adjustments with respect to negative estimates. First, note that the effective spread size in corn futures is about 0.08% or 0.27 cents per bushel. This is a region of "true" spread value at which positive bias in HL and CHL averages over positive estimates is very large, as we argued earlier. As expected, the magnitude of HL and CHL averages of positive-only estimates is larger than the true spread, 0.98% and 1.40%, respectively. This implies estimates about tenfold larger than the effective spreads. Furthermore, the frequency of negative spread estimates is roughly 41% for the high-low estimator and 51% of close-high-low spread estimates, which are of very similar magnitude compared to the data simulated under a true spread value of 0.1%. This provides empirical confirmation of the earlier simulation exercises.

If the researcher had no ability to observe the effective spread, and decided to apply HL and CHL as proxies for trading costs in corn futures, the estimated spreads would bear no relationship with the latent spread, as indicated by the absence of correlation between the spread proxies and true BAS. With our framework, could the researcher infer the poor proxying property of the estimated spread series? In liquid markets such as corn futures, yes. The high frequency of negative estimates suggests that the true, unobserved spread is likely to be small, and small underlying spreads induce upward bias in the positive-only average sample. The proportion of negative estimates is informative of the proxying quality of HL and CHL. Since both daily volatility of the efficient price and true spread level vary in actual data, days with lower volatility can generate less disperse point-estimates, but ultimately since the true spread matters more to define negative estimates, the frequency of negative estimates should align well empirically with the patterns displayed in Figure (III).

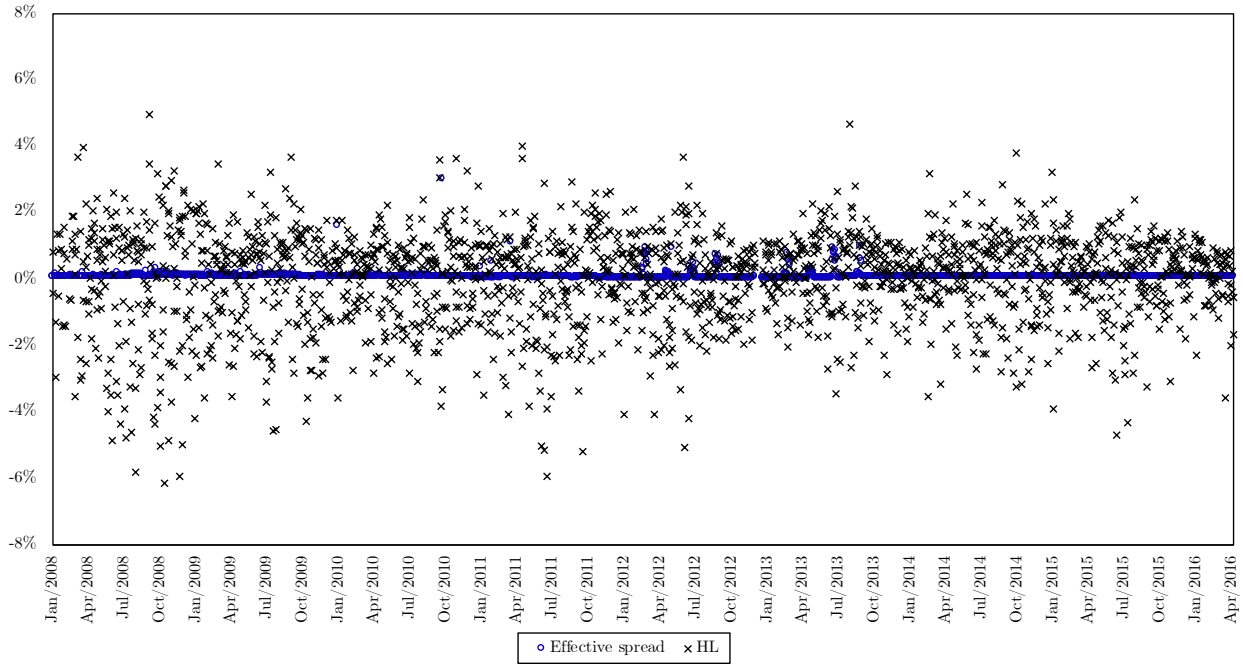
To conclude, in Figure (IX) we show the dispersion of all HL point-estimates and the effective spreads obtained from TAQ data. During 2008, when daily volatility was about 50% higher than the full sample volatility, we can see a greater dispersion of daily estimated

TABLE II
Estimates and Actual Trading Costs in Corn Futures

	Average	Min	Max	Std. Dev.	Correlation
Effective Spread (%)	0.078	0.040	2.991	0.091	
BAS	0.272	0.251	0.525	0.022	
Volatility	3.027	0.399	22.082	2.203	
HL (%)					
Negative adjustment	0.574	0.000	4.939	0.719	−0.017
Full sample	−0.003	−6.158	4.939	1.481	−0.096
Only positives	0.975	0.001	4.939	0.698	0.12***
Frequency negative (%)	41.11				
CHL (%)					
Negative adjustment	0.682	0.000	7.288	1.013	−0.001
Full sample	−0.036	−6.337	7.288	1.760	−0.109
Only positives	1.385	0.013	7.288	1.053	0.21***
Frequency negative (%)	50.80				
Number of days	2041				

Notes: The TAQ data comes from the CME Group BBO dataset. The first sample spans 1/14/2008-12/30/2011. The second dataset starts on 12/2/2013 and ends 4/19/2016. Prices used are always with respect to the nearby corn futures contract, with rollover on the first trading day of the expiration month. The effective spread is computed as $e_i \equiv 2|P_i - M_i|/M_i$, where P_i and M_i are the trade price and the midpoint (arithmetic average) of the outstanding bid-ask spread, respectively, at second i. A daily effective spread is a trade-weighted average of e over all seconds in a given trading day. The variable BAS is the daily simple-average bid-ask spread in cents per bushel. The volatility is measured as the intraday standard deviation of spread midpoints M_i, as in Wang et al. (2014), also in cents/bushel. The high-low and close-high-low spread estimates encompass the same samples used in the high-frequency benchmarks, with rollover on the first-day of the expiration month. In addition to the alternative negative spread adjustments, we implement the overnight and infrequent trading *ad hoc* adjustments from Corwin and Schultz (2012) in both spread proxies for consistency.

spreads, while the effective spread stood around 0.08%. Average estimates after imputing zeros yield $\overline{\text{HL}} = 0.76\%$, but the frequency of negative spreads remained the same as in the full sample.[12] The bias predicted by the proportion of negatives aligns better with the one found in the simulated environment as a consequence of a higher volatility period.



Notes: The figure shows high-low spread point-estimates and effective spreads obtained from daily and TAQ data, respectively, for corn futures prices. There are 2041 daily observations in total.

FIGURE IX
Daily Effective, HL, and CHL Spreads of Corn Futures

[12] CHL daily estimates not reported in Figure (IX) display similar patterns compared to the high-low spreads, although with greater dispersion.

6 Conclusions

Simple bid-ask spread estimators are widely used and actively developed in academic research. While high frequency data enable researchers to observe true spreads and thus directly evaluate the performance of simple estimators, little progress has been made in explaining the misbehavior of these measures. Top-performing estimators often yield point-estimates without economic meaning, being either negative or indeterminate. These are usually attributed to model assumption failures, with a common practice of replacing negative spread estimates with zeros. This overlooks two fundamental problems. First, the frequency of negative estimates might be indicative of systematic misbehavior. Second, the fact that "fixing" negative spreads with zeros increases proxy quality is inconsistent with unbiased positive estimates.

In this paper, we show for the first time that simple bid-ask spread estimators suffer from systematic misbehavior when applied to reasonably liquid markets. Liquid markets provide nearly ideal conditions where these proxies should perform optimally. We demonstrate that in reality, in markets with spreads smaller than 1%, two of the best spread proxies - the high-low and close-high-low estimators - are severely upward-biased and poorly correlated with the true spread. Our results suggest that when 40% or more of spread estimates are negative, this indicates these proxies are inadequate to the application context. The underlying liquidity level we study represents at least half of all stocks traded in the US and the world and most financial and commodity futures markets.

The empirical usefulness of our findings relies on two facts that we establish. First, we derive parsimonious conditions showing that smaller spreads cause the HL and CHL estimators to return negative daily estimates more frequently. Second, positive spread estimates suffer from a larger positive bias when the true spread is small. We then connect these two findings to show that a large frequency of negative spreads identifies positive bias in average spreads. A direct benefit from this approach is that by focusing on the underlying spread magnitude rather than on specific market determinants, our framework predicts proxy misbehavior in any empirical context where trading costs are relatively small, regardless of the asset type.

References

Abdi, Farshid and Angelo Ranaldo (2017) "A simple estimation of bid-ask spreads from daily close, high, and low prices," *The Review of Financial Studies*, Vol. 30, No. 12, pp. 4437–4480.

Adams, Zeno and Thorsten Glück (2015) "Financialization in commodity markets: A passing trend or the new normal?" *Journal of Banking & Finance*, Vol. 60, pp. 93 – 111.

Alizadeh, Sassan, Michael W. Brandt, and Francis X. Diebold (2002) "Range-Based Estimation of Stochastic Volatility Models," *The Journal of Finance*, Vol. 57, No. 3, pp. 1047–1091.

Bleaney, Michael and Zhiyong Li (2015) "The performance of bid-ask spread estimators under less than ideal conditions," *Studies in Economics and Finance*, Vol. 32, No. 1, pp. 98–127.

Brogaard, Jonathan, Dan Li, and Ying Xia (2017) "Stock liquidity and default risk," *Journal of Financial Economics*, Vol. 124, No. 3, pp. 486 – 502.

Chakravarty, Sugato and Asani Sarkar (2003) "Trading Costs in Three U.S. Bond Markets," *The Journal of Fixed Income*, Vol. 13, No. 1, pp. 39–48.

Chen, Xiaohong, Oliver Linton, and Yanping Yi (2017) "Semiparametric identification of the bid-ask spread in extended Roll models," *Journal of Econometrics*, Vol. 200, No. 2, pp. 312 – 325.

Christensen, Kim and Mark Podolskij (2007) "Realized range-based estimation of integrated variance," *Journal of Econometrics*, Vol. 141, No. 2, pp. 323 – 349.

Clark-Joseph, Adam (2013) "Exploratory Trading," *Working Paper*.

Corwin, Shane A. and Paul Schultz (2012) "A simple way to estimate bid-ask spreads from daily high and low prices," *The Journal of Finance*, Vol. 67, No. 2, pp. 719–760.

Demsetz, Harold (1968) "The cost of transacting," *The Quarterly Journal of Economics*, Vol. 82, No. 1, pp. 33–53.

Easley, David, Marcos Lopez de Prado, and Maureen O'Hara (2016) "Discerning information from trade data," *Journal of Financial Economics*, Vol. 120, No. 2, pp. 269 – 285.

Fong, Kingsley Y. L., Craig W. Holden, and Charles A. Trzcinka (2017) "What are the best liquidity proxies for global research?" *Review of Finance*, Vol. 21, No. 1.

French, Kenneth R. and Richard Roll (1986) "Stock return variances," *Journal of Financial Economics*, Vol. 17, No. 1, pp. 5 – 26.

Garman, Mark B. and Michael J. Klass (1980) "On the estimation of security price volatilities from historical data," *The Journal of Business*, Vol. 53, No. 1, pp. 67–78.

Goyenko, Ruslan Y., Craig W. Holden, and Charles A. Trzcinka (2009) "Do liquidity measures measure liquidity?" *Journal of Financial Economics*, Vol. 92, No. 2, pp. 153 – 181.

Harris, Lawrence (1990) "Statistical properties of the Roll serial covariance bid/ask spread estimator," *The Journal of Finance*, Vol. 45, No. 2, pp. 579–590.

Hasbrouck, Joel (2004) "Liquidity in the futures pits: Inferring market dynamics from incomplete data," *The Journal of Financial and Quantitative Analysis*, Vol. 39, No. 2, pp. 305–326.

——— (2009) "Trading costs and returns for U.S. equities: Estimating effective costs from daily data," *The Journal of Finance*, Vol. 64, No. 3, pp. 1445–1477.

Karnaukh, Nina, Angelo Ranaldo, and Paul Söderlind (2015) "Understanding FX Liquidity," *The Review of Financial Studies*, Vol. 28, No. 11, pp. 3073–3108.

Lesmond, David A., Joseph P. Ogden, and Charles A. Trzcinka (1999) "A new estimate of transaction costs," *The Review of Financial Studies*, Vol. 12, No. 5, pp. 1113–1141.

Lin, Chien-Chih (2014) "Estimation accuracy of high-low spread estimator," *Finance Research Letters*, Vol. 11, No. 1, pp. 54 – 62.

Locke, Peter R. and P. C. Venkatesh (1997) "Futures market transaction costs," *Journal of Futures Markets*, Vol. 17, No. 2, pp. 229–245.

Lou, Xiaoxia and Tao Shu (2017) "Price impact or trading volume: Why is the Amihud (2002) measure priced?" *The Review of Financial Studies*, Vol. 30, No. 12, pp. 4481–4520.

Marshall, Ben R., Nhut H. Nguyen, and Nuttawat Visaltanachoti (2011) "Commodity liquidity measurement and transaction costs," *Review of Financial Studies*.

——— (2015) "Frontier market transaction costs and diversification," *Journal of Financial Markets*, Vol. 24, pp. 1 – 24.

McLean, R. David and Jeffrey Pontiff (2016) "Does academic research destroy stock return predictability?" *The Journal of Finance*, Vol. 71, No. 1, pp. 5–32.

Menkveld, Albert J. (2016) "The economics of high-frequency trading: Taking stock," *Annual Review of Financial Economics*, Vol. 8, No. 1, pp. 1–24.

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

O'Hara, Maureen (2015) "High frequency market microstructure," *Journal of Financial Economics*, Vol. 116, No. 2, pp. 257 – 270.

Pagano, Marco and Ailsa Roell (1996) "Transparency and liquidity: A comparison of auction and dealer markets with informed trading," *The Journal of Finance*, Vol. 51, No. 2, pp. 579–611.

Parkinson, Michael (1980) "The extreme value method for estimating the variance of the rate of return," *The Journal of Business*, Vol. 53, No. 1, pp. 61–65.

Pirrong, Craig (1996) "Market liquidity and depth on computerized and open outcry trading systems: A comparison of DTB and LIFFE bund contracts," *Journal of Futures Markets*, Vol. 16, No. 5, pp. 519–543.

Roll, Richard (1984) "A simple implicit measure of the effective bid-ask spread in an efficient market," *The Journal of Finance*, Vol. 39, No. 4, pp. 1127–1139.

Schestag, Raphael, Philipp Schuster, and Marliese Uhrig-Homburg (2016) "Measuring liquidity in bond markets," *The Review of Financial Studies*, Vol. 29, No. 5, pp. 1170–1219.

Thompson, Sarahelen R. and Mark L. Waller (1987) "The execution cost of trading in commodity futures Markets," *Food Research Institute Studies*, Vol. 20, No. 2, pp. 141. – 24.

Tse, Yiuman and Tatyana V. Zabotina (2001) "Transaction costs and market quality: Open outcry versus electronic trading," *Journal of Futures Markets*, Vol. 21, No. 8, pp. 713–735.

Wang, Xiaoyang, Philip Garcia, and Scott H. Irwin (2014) "The behavior of bid-ask spreads in the electronically-traded corn futures market," *American Journal of Agricultural Economics*, Vol. 96, No. 2, pp. 557–577.

Appendix

A Proofs and additional results

Simplified Form of α. Let α be a function of β and γ without explicitly considering each parameter's own arguments (observed high and low prices). The formula for α derived by Corwin and Schultz (2012) is:

$$\alpha\left(\beta,\gamma\right)=\frac{\sqrt{2\beta}-\sqrt{\beta}}{3-2\sqrt{2}}-\sqrt{\frac{\gamma}{3-2\sqrt{2}}}.$$

Call $\varphi\equiv 3-2\sqrt{2}$, $\varphi>0$. After rewriting the above, we have

$$\alpha=\frac{\sqrt{\varphi}\left[\sqrt{\beta}\left(\sqrt{2}-1\right)-\sqrt{\gamma}\sqrt{\varphi}\right]}{\varphi\sqrt{\varphi}}=\frac{\sqrt{\beta}\left(\sqrt{2}-1\right)-\sqrt{\gamma}\left(\sqrt{3-2\sqrt{2}}\right)}{3-2\sqrt{2}}$$

which can be further simplified into

$$\alpha=\frac{\left(\sqrt{2}-1\right)\left(\sqrt{\beta}-\sqrt{\gamma}\right)}{3-2\sqrt{2}}$$

since $\sqrt{3-2\sqrt{2}}=\sqrt{2}-1$. After employing a similar replacement for ψ, we arrive at the simplified version of α used in the main text:

$$\alpha=\frac{\left(3+2\sqrt{2}\right)\left(\sqrt{2}-1\right)\left(\sqrt{\beta}-\sqrt{\gamma}\right)}{\left(3+2\sqrt{2}\right)\left(3-2\sqrt{2}\right)}=\left(\sqrt{2}+1\right)\left(\sqrt{\beta}-\sqrt{\gamma}\right).$$

Simplified Form of the High-Low Estimator. The original closed-form of the high-low spread estimator is

$$S=2\left(\frac{e^{\alpha}-1}{e^{\alpha}+1}\right).$$

A well-known representation of the hyperbolic tangent function $\tanh(\cdot)$ is given by

$$\tanh x \equiv \frac{e^{2x}-1}{e^{2x}+1}$$

for which we can set $x \equiv \alpha/2$, and the simplified version of the HL estimator follows

$$S = 2\tanh\left(\frac{\alpha}{2}\right).$$

Proof of Lemma 1. The range of the high-low spread estimator $S(\alpha)$ only contains negative values when S is evaluated at negative values of α. Hence, negativity of α defines non-positive spread estimates. The parameter α is expressed as $\alpha = (1+\sqrt{2})(\sqrt{\beta}-\sqrt{\gamma})$, which implies that $\gamma > \beta$ determines when $\alpha < 0$. The parameter β always maps each day's range onto $\mathbb{R}_+$. Since γ includes extreme-valued functions, it may take the following values: $\gamma = r_t^2$, $\gamma = r_{t+1}^2$, $\gamma = (h_t - l_{t+1})^2$, or $\gamma = (h_{t+1} - l_t)^2$. In the first two cases, β is always greater than γ. For $\gamma = (h_{t+1} - l_t)^2$, we can write $\gamma > \beta$ as:

$$\underbrace{h_{t+1} - \frac{r_{t+1}}{2}}_{\eta_{t+1}} - \underbrace{\left(\frac{r_t}{2} + l_t\right)}_{\eta_t} + \frac{r_t}{2} + \frac{r_{t+1}}{2} > \sqrt{\beta} \tag{17}$$

and similarly, for $\gamma = (h_t - l_{t+1})^2$:

$$\underbrace{h_t - \frac{r_t}{2}}_{\eta_t} - \underbrace{\left(\frac{r_{t+1}}{2} + l_{t+1}\right)}_{\eta_{t+1}} + \frac{r_t}{2} + \frac{r_{t+1}}{2} > \sqrt{\beta} \tag{18}$$

so that (17) and (18) combined can be stated as:

$$|\eta_{t+1} - \eta_t| > \sqrt{\beta} - \left(\frac{r_t + r_{t+1}}{2}\right) \tag{19}$$

Note that the right-hand side of the negativity condition is always positive, since

$$\left(r_t^2 - r_{t+1}^2\right)^2 + 2\left(r_t^2 + r_{t+1}^2\right) > 0$$

which completes the proof of the lemma.

Detailed Proof of Proposition 1. Begin with the negativity condition of the high-low spread estimator written as:

$$\underbrace{|\eta_{t+1} - \eta_t|}_{f(\eta_t, \eta_{t+1})} > \underbrace{\sqrt{r_t^2 + r_{t+1}^2} - 0.5\,(r_t + r_{t+1})}_{g(r_t, r_{t+1})}. \tag{20}$$

Substituting κ for r_{t+1}/r_t in the above yields

$$|\eta_{t+1} - \eta_t| > r_t\sqrt{1+\kappa^2} - \frac{r_t}{2}(1+\kappa) \tag{21}$$

which after squaring both sides becomes:

$$(\eta_{t+1} - \eta_t)^2 > r_t^2\left(1+\kappa^2\right) - r_t^2(1+\kappa)\sqrt{1+\kappa^2} + \frac{r_t^2}{4}(1+\kappa)^2. \tag{22}$$

Further simplifications return

$$(\eta_{t+1} - \eta_t)^2 > r_t^2\left[\left(1+\kappa^2\right) - (1+\kappa)\sqrt{1+\kappa^2} + \frac{1}{4}(1+\kappa)^2\right] \tag{23}$$

and finally

$$(\eta_{t+1} - \eta_t)^2 > \frac{1}{4}r_t^2\left[(\kappa+1) - 2\sqrt{1+\kappa^2}\right]^2. \tag{24}$$

Without setting $\kappa = 1$, we can replace $r_t = \mathcal{R}_t + S$ in the above and manipulate it as:

$$(\eta_{t+1} - \eta_t)^2 > \mathcal{R}_t^2\Psi(\kappa) + 2\mathcal{R}_t S\Psi(\kappa) + S^2\Psi(\kappa) \tag{25}$$

where $\Psi(\kappa) \equiv \frac{1}{4}\left(\kappa - 2\sqrt{\kappa^2+1}+1\right)^2$. We can then take expectations of both sides and substitute $E[(\eta_{t+1}-\eta_t)^2] \approx 0.61\sigma^2$, $E[\mathcal{R}_t^2] = 4\ln 2\sigma^2$ and $E[\mathcal{R}_t] = (\sqrt{8/\pi})\sigma$:

$$0.61\sigma^2 > \sigma^2\Psi(\kappa)4\ln 2 + 2S\Psi(\kappa)\left(\sqrt{\frac{8}{\pi}}\right)\sqrt{\sigma^2} + \Psi(\kappa)S^2. \tag{26}$$

We assume $\overline{\kappa} = 1$ to make (26) more tractable. For $\kappa \in [0.8, 1.2]$, the function $\Psi(\kappa)$ varies from 0.15 to 0.21, with $\Psi(1) = 0.17$. After plugging 0.17 into the expression above we have

$$0.61\sigma^2 > 0.17\left(2.77\sigma^2 + 3.19S\sqrt{\sigma^2} + S^2\right) \tag{27}$$

which clearly yields the first-order derivatives shown in the proposition.

Proof of Proposition 2. The inequality condition of CHL in (13) expressed in terms of efficient prices is given by

$$(\eta_{t+1}-\eta_t)^2 > \left[\left(q_t\frac{S}{2} + \frac{\mathcal{C}_t}{2} - \frac{\mathcal{H}_t}{2}\right) + \left(\frac{\mathcal{C}_t}{2} - \frac{\mathcal{L}_t}{2}\right)\right]^2 \tag{28}$$

With some algebra, one can transform the above into

$$E\left[(\eta_{t+1}-\eta_t)^2\right] > \frac{S^2}{4} + \frac{1}{4}E\left[(\mathcal{C}_t-\mathcal{H}_t)^2\right] + \frac{1}{2}\left(E\left[\mathcal{C}_t\mathcal{L}_t\right] - E\left[\mathcal{C}_t\mathcal{H}_t\right] + E\left[\mathcal{H}_t\mathcal{L}_t\right]\right) + \frac{1}{4}E\left[(\mathcal{C}_t-\mathcal{L}_t)^2\right] \tag{29}$$

The moments on the right-hand side of (29) are provided in the generating function from Garman and Klass (1980). After performing the pertinent substitutions, we finally have:

$$\left(2 - \frac{k_2}{2}\right)\sigma^2 > \frac{S^2}{4} + \left(1 - \frac{k_2}{4}\right)\sigma^2 \tag{30}$$

for which the relationships stated in the proposition are clearly valid. The case when the CHL spread is negative if $c_t \in (\eta_{t+1}, \eta_t)$ is redundant, therefore we omit it. This completes the proof.

Detailed steps. The close-high-low spread estimator is negatively defined if the following holds:

$$\eta_{t+1} > c_t > \eta_t \tag{31}$$

or, equivalently,

$$\eta_{t+1} - \eta_t > c_t - \eta_t > 0. \tag{32}$$

We can replace observed mid-ranges and close prices with true values:

$$\eta_{t+1} - \eta_t > \mathcal{C}_t + q_t \frac{S}{2} - \left(\frac{\mathcal{H}_t + \mathcal{L}_t}{2} \right) \tag{33}$$

and square both sides such that

$$\left(\eta_{t+1} - \eta_t\right)^2 > \left[\left(q_t \frac{S}{2} + \frac{\mathcal{C}_t}{2} - \frac{\mathcal{H}_t}{2} \right) + \left(\frac{\mathcal{C}_t}{2} - \frac{\mathcal{L}_t}{2} \right) \right]^2 \tag{34}$$

and further simplify it as

$$\begin{aligned} \left(\eta_{t+1} - \eta_t\right)^2 > & \frac{1}{4} \left[q_t^2 S^2 + 2 q_t S \left(\mathcal{C}_t - \mathcal{H}_t \right) + \left(\mathcal{C}_t - \mathcal{H}_t \right)^2 \right] + \\ & + \frac{1}{2} \left[q_t S \mathcal{C}_t + \mathcal{C}_t^2 - q_t S \mathcal{L}_t + \mathcal{C}_t \mathcal{L}_t - \mathcal{C}_t \mathcal{H}_t + \mathcal{H}_t \mathcal{L}_t \right] + \frac{1}{4} \left(\mathcal{C}_t - \mathcal{L}_t \right)^2 . \end{aligned}$$

After taking expectations of the above, we arrive at the intermediate expression in (29), which can be easily solved for with the moments given in Garman and Klass (1980). The final inequality is

$$\left(2 - 2\ln 2\right) \sigma^2 > \frac{S^2}{4} + \frac{\sigma^2}{4} + \frac{1}{2} \left(1 - 2\ln 2\right) \sigma^2 + \frac{\sigma^2}{4} \tag{35}$$

and therefore:

$$(2 - 2\ln 2)\,\sigma^2 > \frac{S^2}{4} + (1 - \ln 2)\,\sigma^2. \tag{36}$$

PROPOSITION 3. *The probability that c_t falls in the interval (η_t, η_{t+1}), and therefore the negativity condition for* CHL *is attained, increases as the spread size decreases.*

PROOF. The probability that the negativity condition of CHL is attained, $\Pr[\eta_t < c_t < \eta_{t+1}]$, is given by

$$\int_{\eta_t}^{\eta_{t+1}} f(c_t)\,dc_t = \int_{l_t}^{h_t} f(c_t)\,dc_t - \underbrace{\left(\int_{\eta_{t+1}}^{h_t} f(c_t)\,dc_t + \int_{l_t}^{\eta_t} f(c_t)\,dc_t\right)}_{\mathcal{K}}.$$

Let $h_t^* = h_t - \delta$ and $l_t^* = l_t + \delta$ represent modified daily high and low prices from a decrease of 2δ in the spread S. Let $\Pr\left[\eta_t^* < c_t < \eta_{t+1}^*\right]$ denote the probability that c_t falls within the negativity interval *ex post* the spread decrease. Then, it follows that

$$\left(\int_{\eta_{t+1}}^{h_t} f(c_t)\,dc_t + \int_{l_t}^{\eta_t} f(c_t)\,dc_t\right) \geq \left(\int_{\eta_{t+1}}^{h_t^*} f(c_t)\,dc_t + \int_{l_t^*}^{\eta_t} f(c_t)\,dc_t\right)$$

since

$$\left(\int_{\eta_{t+1}}^{h_t^*} f(c_t)\,dc_t + \int_{l_t^*}^{\eta_t} f(c_t)\,dc_t\right) \leq \left(\int_{h_t-\delta}^{h_t} f(c_t)\,dc_t + \int_{\eta_{t+1}}^{h_t-\delta} f(c_t)\,dc_t + \int_{l_t+\delta}^{\eta_t} f(c_t)\,dc_t + \int_{l_t}^{l_t+\delta} f(c_t)\,dc_t\right)$$

and therefore $\int_{\eta_t}^{\eta_{t+1}} f(c_t)\,dc_t \leq \int_{\eta_t^*}^{\eta_{t+1}^*} f(c_t)\,dc_t$. The equality holds if and only if $\mathcal{K} = 0$. This can be the case only when $\eta_t = \eta_{t+1}$. Hence, $\Pr\left[\eta_t^* < c_t < \eta_{t+1}^*\right] > \Pr[\eta_t < c_t < \eta_{t+1}]$ and we conclude the proof.

B Generality of Lemma 1

In Lemma 1, we argue that the negativity condition defined as $|\eta_{t+1} - \eta_t| > \sqrt{\beta} - r$ is necessary and sufficient for obtaining negative-only HL estimates. Throughout we assumed no consecutive-day price ties and well-defined daily ranges. Here, we relax the first assumption and show that any price tie yields strictly positive spread values.

Identical consecutive-day low prices

Let $h_t > h_{t+1}$ (w.l.o.g.) and $l_t = l_{t+1} \equiv a$. Given the definitions for β and γ, and the negativity condition slightly restated as $\gamma \geq \beta$, we have:

$$(h_t - a)^2 \geq (h_t - a)^2 + (h_{t+1} - a)^2 \tag{37}$$

which cannot hold since $h_{t+1} > a$.

Identical consecutive-day high prices

Consider when $h_t = h_{t+1} = b$ and $l_t < l_{t+1}$ (w.l.o.g.). The negativity conditions becomes:

$$(b - l_t)^2 \geq (b - l_t)^2 + (b - l_{t+1})^2 \tag{38}$$

which is unfeasible since $b > l_{t+1}$.

Identical consecutive-day low and high prices

Now, let $l_t = l_{t+1} = a$ and $h_t = h_{t+1} = b$. It is straightforward to see that $\gamma = 2\beta$, $\beta > 0$. Therefore, the spread is always strictly positive.

C Calibration of κ

The constant κ ultimately depends on S and σ^2, since it is defined as the ratio of consecutive-day observed ranges. As an alternative to arbitrarily assume $\overline{\kappa} = 1$, κ could be calibrated from the data as $\overline{\kappa} = I^{-1} \sum_{i=1}^{I} \overline{\kappa}_i$, where $\overline{\kappa}_i$ is the average value of the constant

estimated for financial product i:

$$\overline{\kappa}_i = T^{-1} \sum_{t=1}^{T-1} \frac{r_{i,t+1}}{r_{i,t}}. \tag{39}$$

To evaluate how well our assumed $\overline{\kappa}$ represents actual data, we obtain end-of-day high and low prices of S&P 500 stocks from 2000 to 2018 to estimate stock-specific values of $\overline{\kappa}_i$. These estimates are reported in Table (A.1). The average $\overline{\kappa}$ of 1.1 calibrated from the S&P 500 stock data is very close to the unity, which indicates that our simplifying assumption is fairly reasonable. Since Ψ is increasing in the neighborhood of 1, higher $\overline{\kappa}$ values only scale up $E[g]$, so that the comparative statics in Proposition 1 remains unchanged.

TABLE A.1
ESTIMATES OF κ FOR S&P 500 STOCKS

Ticker	$\overline{\kappa}$	Days	Ticker	$\overline{\kappa}$	Days	Ticker	$\overline{\kappa}$	Days	Ticker	$\overline{\kappa}$	Days	Ticker	$\overline{\kappa}$	Days
A	1.14	4632	CLX	1.13	4632	GD	1.14	4632	MAC	1.14	4632	RL	1.15	4632
AAL	1.14	3191	CMA	1.13	4632	GE	1.13	4632	MAR	1.14	4632	RMD	1.17	4632
AAP	1.16	4154	CMCSA	1.12	4632	GGP	1.17	4632	MAS	1.13	4632	ROK	1.14	4632
AAPL	1.14	4632	CME	1.15	3897	GILD	1.12	4632	MAT	1.14	4632	ROP	1.16	4632
ABBV	1.15	1363	CMG	1.13	3108	GIS	1.12	4632	MCD	1.13	4632	ROST	1.14	4632
ABC	1.16	4632	CMI	1.14	4632	GLW	1.13	4632	MCHP	1.12	4632	RSG	1.16	96
ABMD	1.18	4632	CMS	1.12	4632	GM	1.14	1895	MCK	1.15	4632	RTN	1.14	4632
ABT	1.13	4632	CNC	1.18	4144	GOOG	1.13	3470	MCO	1.16	4632	SBAC	1.16	4632
ACN	1.14	4243	CNP	1.14	4632	GOOGL	1.13	3470	MDLZ	1.14	4268	SBUX	1.12	4632
ADBE	1.12	4632	COF	1.13	4632	GPC	1.13	4632	MDT	1.13	4632	SCG	1.13	4632
ADI	1.12	4632	COG	1.13	4632	GPN	1.16	4371	MET	1.12	4567	SCHW	1.14	4632
ADM	1.13	4632	COL	1.14	4266	GPS	1.12	96	MGM	1.15	4632	SEE	1.15	4632
ADP	1.13	4632	COO	1.17	4632	GRMN	1.18	4395	MHK	1.17	4632	SHW	1.14	4632
ADS	1.19	4266	COP	1.11	4632	GS	1.12	4632	MKC	1.15	4632	SIVB	1.14	4632
ADSK	1.14	4632	COST	1.12	4632	GT	1.15	4632	MLM	1.16	4632	SJM	1.17	4632
AEE	1.11	96	COTY	1.15	1251	GWW	1.15	4632	MMC	1.15	4632	SLB	1.11	4632
AEP	1.12	4632	CPB	1.13	4632	HAL	1.12	4632	MMM	1.13	4632	SLG	1.15	4632
AES	1.14	4632	CPRT	1.15	4632	HAS	1.15	4632	MNST	1.25	4632	SNA	1.14	4632
AET	1.15	4632	CRM	1.15	3510	HBAN	1.12	4632	MO	1.14	4632	SNPS	1.14	4632
AFL	1.13	4632	CSCO	1.14	4632	HBI	1.15	2954	MOS	1.14	4632	SO	1.11	4632
AGN	1.15	4632	CSX	1.13	4632	HCA	1.16	1819	MPC	1.12	1745	SPG	1.13	4632
AIG	1.14	4632	CTAS	1.13	4632	HCP	1.14	4632	MRK	1.13	4632	SPGI	1.14	4632
AIV	1.15	4632	CTL	1.15	4632	HD	1.12	4632	MRO	1.11	4632	SRCL	1.16	4632
AIZ	1.14	3605	CTSH	1.15	4632	HES	1.12	4632	MS	1.12	4632	SRE	1.13	4632

AJG	1.14	4632	CTXS	1.13	4632	HFC	1.17	4632	MSCI	1.16	2653	STI	1.12	4632
AKAM	1.14	4632	CVS	1.14	4632	HIG	1.13	4632	MSFT	1.12	4632	STT	1.14	4632
ALB	1.16	4632	CVX	1.11	4632	HII	1.15	1811	MSI	1.14	4632	STX	1.15	3894
ALGN	1.17	4361	CXO	1.12	2726	HLT	1.14	1124	MTB	1.15	4632	STZ	1.16	4632
ALK	1.14	4632	D	1.12	4632	HOG	1.15	4632	MTD	1.17	4632	SWK	1.14	4632
ALL	1.13	4632	DAL	1.13	2790	HOLX	1.16	4632	MU	1.12	4632	SWKS	1.13	4632
ALLE	1.14	1141	DE	1.13	4632	HON	1.12	4632	MYL	1.16	4632	SYF	1.15	966
ALXN	1.16	4632	DFS	1.12	2761	HP	1.12	4632	NBL	1.12	4632	SYK	1.16	4632
AMAT	1.12	4632	DG	1.16	2150	HPE	1.15	659	NCLH	1.15	1351	SYMC	1.14	4632
AMD	1.14	4632	DGX	1.14	4632	HPQ	1.13	4632	NDAQ	1.19	4008	SYY	1.14	4632
AME	1.15	4632	DHI	1.13	4632	HRB	1.16	4632	NEE	1.12	4632	T	1.12	4632
AMG	1.15	4632	DHR	1.13	4632	HRL	1.14	4632	NEM	1.13	4632	TAP	1.17	4632
AMGN	1.12	4632	DIS	1.12	4632	HRS	1.16	4632	NFLX	1.15	4034	TDG	1.17	3075
AMP	1.13	3199	DISCA	1.15	3247	HSIC	1.15	4632	NFX	1.14	4632	TEL	1.15	2761
AMT	1.15	4632	DISCK	1.14	2442	HST	1.13	4632	NI	1.12	4632	TGT	1.13	4632
AMZN	1.13	4632	DISH	1.16	4632	HSY	1.14	4632	NKE	1.14	4632	TIF	1.14	4632
ANDV	1.14	4632	DLR	1.16	3420	HUM	1.16	4632	NKTR	1.16	4632	TJX	1.13	4632
ANSS	1.17	4632	DLTR	1.14	4632	IBM	1.12	4632	NLSN	1.15	1848	TMK	1.12	4632
ANTM	1.15	4175	DOV	1.14	4632	ICE	1.13	3155	NOC	1.14	4632	TMO	1.13	95
AON	1.15	4632	DRE	1.14	4632	IDXX	1.16	4632	NOV	1.13	4632	TPR	1.15	4439
AOS	1.18	4632	DRI	1.14	4632	IFF	1.14	4632	NRG	1.16	3649	TRIP	1.15	1630
APA	1.11	4632	DTE	1.12	4632	ILMN	1.19	4488	NSC	1.13	4632	TROW	1.13	4632
APC	1.12	4632	DUK	1.12	4632	INCY	1.15	4632	NTAP	1.13	4632	TRV	1.13	4632
APD	1.13	4632	DVA	1.18	4632	INFO	1.15	995	NTRS	1.12	4632	TSCO	1.17	4632
APH	1.15	4632	DVN	1.12	4632	INTC	1.11	4632	NUE	1.13	4632	TSN	1.17	4632
APTV	1.15	1643	DWDP	1.13	4632	INTU	1.13	4632	NVDA	1.13	4632	TSS	1.16	4632
ARE	1.17	4632	DXC	1.16	4632	IP	1.12	4632	NWL	1.14	4632	TTWO	1.14	4632
ARNC	1.13	4632	EA	1.13	4632	IPG	1.15	4632	NWS	1.14	1247	TWTR	1.14	1148

ATVI	1.14	4632	EBAY	1.13	4632	IPGP	1.16	2885	NWSA	1.14	1247	TXN	1.12	4632
AVB	1.15	4632	ECL	1.13	4632	IQV	1.15	1275	O	1.13	4632	TXT	1.15	4632
AVGO	1.13	2220	ED	1.11	4632	IR	1.13	4632	OKE	1.14	4632	UA	1.14	735
AVY	1.14	4632	EFX	1.15	4632	IRM	1.18	4632	OMC	1.13	4632	UAA	1.14	3153
AWK	1.12	2545	EIX	1.14	4632	ISRG	1.15	4517	ORCL	1.12	4632	UAL	1.14	3101
AXP	1.13	4632	EL	1.15	4632	IT	1.16	4632	ORLY	1.14	4632	UDR	1.09	95
AZO	1.15	4632	EMN	1.13	4632	ITW	1.13	4632	OXY	1.11	4632	UHS	1.16	4632
BA	1.13	4632	EMR	1.13	4632	IVZ	1.15	4632	PAYX	1.12	4632	ULTA	1.14	2668
BAC	1.12	4632	EOG	1.12	4632	JBHT	1.14	4632	PBCT	1.14	4632	UNH	1.14	4632
BAX	1.14	4632	EQIX	1.16	4478	JCI	1.14	4632	PCAR	1.13	4632	UNM	1.15	4632
BBT	1.12	4632	EQR	1.14	4632	JEC	1.15	4632	PCG	1.13	4632	UNP	1.13	4632
BBY	1.14	4632	EQT	1.14	4632	JEF	1.17	4632	PEG	1.12	4632	UPS	1.13	4632
BDX	1.13	4632	ES	1.13	4632	JNJ	1.12	4632	PEP	1.12	4632	URI	1.17	4632
BEN	1.13	4632	ESRX	1.15	4632	JNPR	1.13	4632	PFE	1.12	4632	USB	1.12	4632
BHF	0.98	2700	ESS	1.15	4632	JPM	1.12	4632	PFG	1.13	4180	UTX	1.13	4632
BHGE	1.11	4632	ETFC	1.14	4632	JWN	1.13	4632	PG	1.13	4632	V	1.13	2569
BIIB	1.12	4632	ETN	1.14	4632	K	1.13	4632	PGR	1.14	4632	VAR	1.17	4632
BK	1.13	4632	ETR	1.13	4632	KEY	1.12	4632	PH	1.14	4632	VFC	1.14	4632
BKNG	1.14	4632	EVHC	1.17	1208	KHC	1.11	733	PHM	1.13	4632	VIAB	1.14	3143
BLK	1.19	4632	EVRG	1.16	5	KIM	1.14	4632	PKG	1.15	4614	VLO	1.13	4632
BLL	1.14	4632	EW	1.16	4574	KLAC	1.12	4632	PKI	1.15	4632	VMC	1.15	4632
BMY	1.14	4632	EXC	1.12	4632	KMB	1.12	4632	PLD	1.15	4632	VNO	1.15	4632
BR	1.15	2812	EXPD	1.14	4632	KMI	1.13	1837	PM	1.13	2571	VRSK	1.16	2177
BSX	1.15	4632	EXPE	1.13	3238	KMX	1.16	4632	PNC	1.13	4632	VRSN	1.13	4632
BWA	1.15	4632	EXR	1.13	3473	KO	1.12	4632	PNR	1.15	4632	VRTX	1.14	4632
BXP	1.14	4632	F	1.14	4632	KORS	1.13	1624	PNW	1.12	4632	VTR	1.16	4632
C	1.12	4632	FAST	1.14	4632	KR	1.14	4632	PPG	1.21	96	VZ	1.12	4632
CA	1.13	4632	FB	1.13	1518	KSS	1.14	4632	PPL	1.10	95	WAT	1.15	4632

CAG	1.13	4632	FBHS	1.14	1687	KSU	1.15	4632	PRGO	1.16	4632	WBA	1.14	4632
CAH	1.14	4632	FCX	1.13	4632	L	1.13	4632	PRU	1.13	4144	WDC	1.14	4632
CAT	1.12	4632	FDX	1.13	4632	LB	1.14	4632	PSA	1.14	4632	WEC	1.12	4632
CB	1.14	4632	FE	1.12	4632	LEG	1.13	4632	PSX	1.12	1544	WELL	1.13	4632
CBOE	1.13	2005	FFIV	1.14	4632	LEN	1.13	4632	PVH	1.16	4632	WFC	1.12	4632
CBRE	1.13	95	FIS	1.17	4263	LH	1.15	4632	PWR	1.16	4632	WHR	1.15	4632
CBS	1.13	3143	FISV	1.13	4632	LKQ	1.15	3689	PX	1.13	4632	WLTW	1.18	4269
CCI	1.16	4632	FITB	1.12	4632	LLL	1.15	4632	PXD	1.12	4632	WM	1.14	4632
CDNS	1.15	4632	FL	1.14	4632	LLY	1.13	4632	PYPL	1.13	733	WMB	1.14	4632
CELG	1.14	4632	FLIR	1.18	4632	LMT	1.14	4632	QCOM	1.12	4632	WMT	1.12	4632
CERN	1.16	4632	FLR	1.15	4400	LNC	1.13	4632	QRVO	1.14	859	WRK	1.16	740
CF	1.13	3223	FLS	1.16	4632	LNT	1.13	4632	RCL	1.14	4632	WU	1.14	2936
CFG	1.12	928	FLT	1.16	1877	LOW	1.13	4632	RE	1.14	4632	WY	1.13	4632
CHD	1.14	4632	FMC	1.14	4632	LRCX	1.12	4632	REG	1.15	4632	WYNN	1.13	3926
CHRW	1.15	4632	FOX	1.15	4632	LUV	1.13	4632	REGN	1.15	4632	XEC	1.13	3948
CHTR	1.19	2116	FOXA	1.15	4632	LYB	1.15	2038	RF	1.13	4632	XEL	1.12	4632
CI	1.15	4632	FRT	1.14	4632	M	1.14	4632	RHI	1.15	4632	XL	1.16	4632
CINF	1.13	4632	FTI	1.14	4266	MA	1.14	3025	RHT	1.14	4632	XLNX	1.12	4632
CL	1.12	4632	FTV	1.14	481	MAA	1.15	4632	RJF	1.13	4632	XOM	1.11	4632
												XRAY	1.15	4632
												XRX	1.15	4632
												XYL	1.14	1668
												YUM	1.14	4632
												ZBH	1.16	4239
												ZION	1.13	4632
												ZTS	1.13	1342

Notes: The constant $\overline{\kappa}$ is the average ratio of consecutive-daily log ranges over 2000-2018 for each S&P 500 stock. For securities that started trading after 2000, we begin the time series as early as data becomes available.