THE HIGH–LOW SPREAD ESTIMATOR IS NOT WELL–BEHAVED
IN COMMODITY MARKETS

BY

PEDRO TREMACOLDI ROSSI

THESIS
Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Agricultural & Applied Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Master’s Committee:
Professor Scott H. Irwin, Chair
Professor Philip Garcia
Professor Michel A. Robe
Associate Professor Maria Teresa Serra Devesa
ABSTRACT

In spite of the increasing availability of high-quality data and the possibility of obtaining direct commodity trading costs in recent periods, historical series of transaction costs still require bid-ask spread estimation. In this work, we verify whether the popular high-low spread estimator performs well in commodity markets so that it can be used to construct long-term cost estimates. We find that the estimator suffers both from measurement error increasing in the volatility-to-spread ratio and consistently positive error in a variety of empirical and experimental settings. As the measurement error in the high-low estimator depends on ex-ante knowledge about the usually unobserved true spread level, we conclude that the spread measure is not well-behaved and should be avoided in commodity markets.
ACKNOWLEDGEMENTS

Nisi credideritis, non intelligetis — Saint Augustine, De libero arbitrio

This thesis is naturally the result of support from many people. From unconditional encouragement to alternative directions when research failed to progress as expected, my advisor, Prof. Scott Irwin, has given fundamental guidance for the completion of this work. I am also very grateful for the generous research assistantship he offered throughout my M.S. program. Special thanks to Professor Philip Garcia who along with Prof. Irwin unveiled a turning point in this research. Tatiana Mocanu, who patiently read multiple earlier drafts to which she has brilliantly contributed, and for always standing by my side, my heartfelt gratitude. I am grateful to Professors Michel Robe and Maria Teresa Serra Devesa for valuable comments and suggestions as Committee members. To my parents and friends for endless support. Finally, to ACE Department’s faculty, staff and graduate students for promoting the best research environment one could ask for. All remaining errors are my own.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>Chapter 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 2 THE CORWIN AND SCHULTZ HIGH-LOW SPREAD ESTIMATOR</td>
<td>6</td>
</tr>
<tr>
<td>2.1 The model</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Empirical shortcomings</td>
<td>10</td>
</tr>
<tr>
<td>Chapter 3 HISTORICAL SPREAD ESTIMATION</td>
<td>14</td>
</tr>
<tr>
<td>3.1 Data</td>
<td>14</td>
</tr>
<tr>
<td>3.2 Transaction costs in commodity markets</td>
<td>14</td>
</tr>
<tr>
<td>Chapter 4 MEASUREMENT ERROR IN THE CSHL ESTIMATOR</td>
<td>21</td>
</tr>
<tr>
<td>4.1 The importance of the volatility-to-spread ratio</td>
<td>21</td>
</tr>
<tr>
<td>4.2 The behavior of $\alpha$, $\beta$ and $\gamma$</td>
<td>26</td>
</tr>
<tr>
<td>4.3 Actual trading costs and high-low spread estimates</td>
<td>34</td>
</tr>
<tr>
<td>Chapter 5 CONCLUSIONS</td>
<td>40</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>42</td>
</tr>
<tr>
<td>APPENDIX A: Simplified form of $\alpha$</td>
<td>46</td>
</tr>
<tr>
<td>APPENDIX B: Strict positivity of minor cases</td>
<td>47</td>
</tr>
</tbody>
</table>
LIST OF TABLES

3.1 Selected commodity futures contracts ........................................ 15
3.2 Commodity futures high-low spread estimates .......................... 17

4.1 Main simulation results ............................................................ 24
4.2 CHSL estimates and actual trading costs in corn futures .............. 38
4.3 Correlation of CSHL estimates and high-frequency benchmark .... 38
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Overnight returns adjustment</td>
<td>12</td>
</tr>
<tr>
<td>2.2</td>
<td>Liquidity adjustment</td>
<td>13</td>
</tr>
<tr>
<td>3.1</td>
<td>Commodity futures historical CSHL spread estimates</td>
<td>20</td>
</tr>
<tr>
<td>4.1</td>
<td>CSHL estimates under almost ideal conditions</td>
<td>25</td>
</tr>
<tr>
<td>4.2</td>
<td>General cases of combinations with daily high and low prices</td>
<td>31</td>
</tr>
<tr>
<td>4.3</td>
<td>Numerical example of Case 2</td>
<td>33</td>
</tr>
<tr>
<td>4.4</td>
<td>CSHL estimator measurement error</td>
<td>36</td>
</tr>
<tr>
<td>4.5</td>
<td>Daily average bid-ask spreads for corn futures: 2008-2011</td>
<td>37</td>
</tr>
<tr>
<td>4.6</td>
<td>Daily average bid-ask spreads for corn futures: 2013-2016</td>
<td>37</td>
</tr>
</tbody>
</table>
Chapter 1

INTRODUCTION

Transaction costs have long been one of the paramount areas in the economic efficiency literature. Lower transaction costs in a marketplace ease economic interactions, contributing to welfare gains as trade expands. The emergence, persistence, and dynamics of transaction costs have been widely scrutinized in empirical applications, particularly in a subset of finance popularized as market microstructure, due to readily available data on recorded transactions of financial assets. Understood as the cost dimension of liquidity, transaction costs are important stable entities to measure liquidity constraints in financial markets, specially in light of recent evidence supporting that even when market efficiency is jeopardized, trading activity might be extremely resilient.\(^1\) In other words, if one understands transaction costs as the cost of participating in a market, and efficiency on normative grounds, issues regarding the latter not necessarily affect the former, whereas high transactions costs might induce liquidity to vanish, which in turn affects the proper functioning of a market.

Following the conceptual framework presented by Demsetz (1968) focused on transacting with immediacy, bid-ask spreads (BAS) have received primary attention as the most important quantifiable element in comprising transaction costs.\(^2\) Since BAS were scarcely observable before the appearance of trade-and-quote (TAQ) data, developing successful bid-ask spread estimators, or more broad liquidity measures, became an active area of research. Arguably, the degree of success of an estimator comprises not only its intrinsic properties and empirical reliability, but how easily implementable and meaningful it might be vis-à-vis specific market structures. Early work established pioneer spread estimators that would require only low frequency data (hence implementability would come as a side benefit), and whose properties were mainly subject only to theoretical analysis (Roll (1984), French and Roll (1986), Thompson and Waller (1987), Harris (1990)). However, both empirical reliability and meaningfulness demanded data quality that was virtually nonexistent until

\(^1\)In the 2005-2010 grain futures convergence “dilemma”, Garcia et al. (2015) point that trading volume doubled in spite of serious problems of futures-cash prices convergence.

\(^2\)Other seminal papers (e.g. (Tinic (1972) and Stoll (1978)) broadened the initial Demsetzian analytical structure of trading costs.
the mid-1990s, which hampered performance evaluation. In addition, most estimators are derived assuming a particular market microstructure, and the consequences arising from changing the original market environment are usually far from obvious without the ability of observing the true spread.

The electronic trading paradigm shift enabled direct observation of actual trading costs, making the computation of high-frequency spread benchmarks possible. To a certain extent, actual liquidity benchmarks provide a criterion for choosing an estimator whenever the actual trading cost is unavailable, preserving the relevance of estimation options in both theoretical and applied work.\textsuperscript{3} There are at least two reasons supporting the interest in this approach. First, inasmuch as TAQ data made observance of market microstructural features yet unprecedented feasible, high frequency data introduced its own drawbacks into empirical analysis (Boehmer et al. (2007), Fabozzi et al. (2011)).\textsuperscript{4} Second, historical series of transaction costs require some sort of estimation for periods without actual information on actual trading costs. Therefore, one may test how different estimator choices perform when actual trading cost information is obtainable, and under the assumption that the performance results extend beyond the tested sample period, construct historical transaction costs.

As the first comprehensive study to confront alternative spread estimators against transaction-based liquidity benchmarks in equity markets by Goyenko et al. (2009) well illustrates, cross-sectional correlation between a spread estimator and a liquidity benchmark provides a straightforward way to compare how satisfactorily different estimators proxy transaction costs. In this context, an appealing spread estimator that has been recently proposed by Corwin and Schultz (2012), showed to perform well in equity markets (Fong et al. (2017)). The high-low spread (CSHL, hereafter) estimator explores the information contained in daily extreme prices in order to derive a low-frequency, percent-spread measure that relies only on daily data. Corwin and Schultz (2012) demonstrate that monthly high-low spreads have an average cross-sectional correlation with effective spreads for NYSE, Amex, and NASDAQ securities of nearly 83% in the period of 1993-2006. In a similar vein, Fong et al. (2017) find the CSHL estimator to be one of the best monthly percent-cost

\textsuperscript{3}In fact, one could expect that the importance of understanding spread estimators would fade away, since using the true variable (the observed spread) should always precede the use of its proxies (low frequency estimators).

\textsuperscript{4}Concerns with respect to data cleansing, synchronization and report errors are likely to increase in trading frequency, ultimately requiring the researcher’s decision over how much precision or robustness is desirable (Brownlees and Gallo (2006)).
proxies when comparing the performance of multiple spread estimators in stock markets in various countries. Moreover, when analyzing the role of liquidity in FX markets, Karnaukh et al. (2015) find that their high-frequency liquidity benchmark is highly correlated with CSHL. Similarly, Schestag et al. (2016) find a superior performance of the estimator in US corporate bond markets. As the high-low measure shows empirical reliability for recent periods, “these results suggest that the high-low spread estimator may be superior to other estimators for historical analyses”, as claimed by Corwin and Schultz (2012).

The attractiveness of CSHL, also due to its easy implementability and reasonable assumptions, justifies its fast popularization as a spread proxy in studies which need to account for transaction costs in a variety of settings. Adams and Glück (2015) employ the high-low estimator to control for the effect of liquidity on risk spillovers in commodities. McLean and Pontiff (2016) use the estimator as the liquidity variable to survey whether investors display learning after academic publications. Further examples include Easley et al. (2016), who apply the high-low measure to account for spread effects when inferring trade intentions from observed data, and the estimation of transaction costs in frontier markets in Marshall et al. (2015).

In lieu with studies favoring the high-low spread measure, Marshall et al. (2011) evaluate a number of estimator choices in commodity markets, without finding an overall good performance of CHSL. Since they do not directly address the causes of poor high-low estimates in commodity markets, it is still uncertain whether the CSHL estimator is a suitable proxy for constructing historical spreads in futures trading. The importance of determining CSHL’s reliability extends beyond the long standing need of historical trading costs in commodity markets, as studies relying on commodity BAS estimation may be subject to measurement error issues stemming from an inappropriate estimator choice.

Some previous results may justify the discrepancy in estimation performance between equity and futures trading. Commodity futures markets offer a different marketplace environment than the one faced by traders in equity markets (Locke and Venkatesh (1997)). At least since Brorsen (1989), who found scalpers costs in the corn futures markets to be smaller than in stocks, commodity markets are understood as a trading framework whose idiosyncratic features create an opportunity for assessing both the meaningfulness and empirical reliability of spread estimators, usually developed under equity trading microstructure.\(^5\) There is already evidence that spread measures commonly applied to stock markets, per-

\(^5\)For an example, see Bryant and Haigh (2002).
form poorly in commodity futures. Locke and Venkatesh (1997) concluded that the popular Roll measure (Roll (1984)) “bear[s] no relationship to futures market transaction costs”. Moreover, historical estimates of transaction costs are generally sparse, and only recently the behavior of bid-ask spreads in commodity markets has been studied more consistently, mainly for markets with some degree of electronic trading (Frank and Garcia (2011) and Wang et al. (2014)). Thus, how spreads have evolved over time in futures trading and how to safely obtain historical transaction costs with estimators developed in different market microstructures are still important research topics.

In this work we take a much narrower performance approach than the cross-sectional multi-estimator comparison framework in Goyenko et al. (2009) and thoroughly analyze the behavior of the high-low spread estimator in the context of commodity trading. The CSHL estimator relies on daily high and low prices as identification sources for buyer- and seller-initiated orders, respectively, and volatility proportionality in a 2-day interval to calculate bid-ask spreads. We estimate historical transaction costs for six commodity futures markets (corn, soybean, wheat, sugar, live cattle, and oil) dating back to the 1950s. The estimates display abnormal trends in trading costs over time, with an overall increasing pattern. Under the suspicion that the great number of negative empirical estimates might be penalizing the general performance of the CSHL estimator, we create a simulation environment to isolate the effect of the true spread size and volatility on high-low spread estimates. We find strong evidence of consistent positive measurement error increasing in the volatility-to-spread ratio, even when volatility and actual spreads are constant across days and market conditions are nearly ideal. We also show that the proportion of negative estimates appear to increase monotonically in the measurement error size, as fewer non-negative observations contribute towards the convergence of the sample average to the true spread level.

We render a formal treatment to the CSHL measure in order to determine the cases where the high-low estimator is negative. Surprisingly, these conditions show that negativity arises in fairly common cases in trading, which amount to nearly 80% of trading days. By employing TAQ data from 2008 to 2016, we assess the proxying quality of CSHL with respect to effective spreads in the corn market. Even when the high-low measure can only be positively defined, there is consistent positive measurement error, and overall correlation between the proxy and the high frequency benchmark indicate a lack of relationship between both. Altogether, we suggest that the high-low spread estimator performs poorly in markets

---

6The derivation of long-term transaction costs estimates that are reliable, would help ascertain whether electronic trading brought efficiency gains to futures trading (Pirrong (1996)).
with a small spread size, such as commodity markets, which makes historical transaction estimates based on the measure unreliable. No averaging process yields better results, since positive-only estimates suffer from positive measurement error. Ultimately the estimator may depend on the size of the true spread so that there are enough non-negative underestimates offsetting overstated spreads. The abnormal behavior and reliance on the unobserved spread make the high-low spread estimator not well-behaved and an unstable option to proxy trading costs in commodity markets.
Chapter 2

THE CORWIN AND SCHULTZ HIGH-LOW SPREAD ESTIMATOR

2.1 The model

Characterize a futures contract by the triplet \((C,M,Y)\), where \(C\) is the reference commodity, and \(M\) and \(Y\) the maturity month and year, respectively.\(^1\) A contract life-time is encompassed by \(t = 1,\ldots, N\) trading days, where our aggregate time quantity of interest is \(\bigcup_{k=1}^{N-1} [t_k, t_{k+1}]\), i.e., pairwise consecutive days. Every trading day \(t\) is spanned by \(\mathcal{T}_t = [0, T] \subset \mathbb{N}\), where \(\mathcal{T}_t\) denotes all discrete points in time \(\tau\) for which price quotes of the reference futures contract are observed. Define the sequence of all prices observed on \(t\) as \(\mathcal{G}_t \equiv \{(P_\tau)_{\tau \in \mathcal{T}_t} : \forall P \in \mathbb{R}_+\}\). Hence, \(\inf\{\mathcal{T}_t : 0 \leq \tau \leq T\}\) corresponds to the opening price \(O_t\) on day \(t\) and \(\sup_{[0, T]} \mathcal{T}_t\) to the daily closing price, \(C_t\). Further, denote \((\tilde{H}^O_t, \tilde{L}^O_t) \equiv (\max\{\mathcal{G}_t\}, \min\{\mathcal{G}_t\})\) as the daily ordered pair of observed high and low prices. Set the difference \(R_t \geq (\tilde{H}_t - \tilde{L}_t)\), where \(\tilde{\gamma} \equiv \ln \gamma\), as the daily high-low ratio and define the two-day ratio as \(R_{t,t+1} \geq (\max\{\tilde{H}_{t+1}, \tilde{H}_t\} - \min\{\tilde{L}_{t+1}, \tilde{L}_t\})\).\(^2\) Throughout the paper, we consider that the equality holds in both daily and two-day log ranges.

The paper by Corwin and Schultz (2012) introduces a powerful and versatile low frequency bid-ask spread estimator using daily high and low prices. CSHL’s major theoretical strength is due to its derivation primarily from two basic assumptions:

**Assumption 1** (Daily high-low ratio as the bid-ask spread). At any given trading day \(t\), \(R_t\) reflects an asset’s fundamental volatility (true variance) and its bid-ask spread, since \(\tilde{H}^O_t\) is a buyer-initiated order and \(\tilde{L}^O_t\) is a seller-initiated order.

This assumption is sustained on the typical tick test mechanism to infer buyer and seller-initiated orders from Lee and Ready (1991). The other assumption is also standard in the finance literature:

\(^1\)This representation mimics the usual commodity-month-year codes.

\(^2\)The ratio \(R\) is also commonly known as log range in the finance literature (Alizadeh et al. (2002)).
Assumption 2 (Spreads are constant over time). Under a geometric Brownian motion, the true variance of an asset is proportional to the length $t = j - 1, j > 1$. For instance, $\sum_{t=1}^{2} R_t$ and $R_{t,t+1}$ reflect the volatility over 2 days in the same manner. However, the spread component of $R_t$ is constant over every two-day interval. That is, when adding two consecutive high-low ratios, one accounts for twice the spread but when obtaining $R_{t,t+1}$ only one spread is measured.\footnote{To illustrate, say $R_1 = 1.3$, $R_2 = 1.7$, and $R_{1,2} = 2.3$. The summation over days 1 and 2, $1.3 + 1.7 = 3$, incorporates the spread on the first day and the spread on the second day, which are assumed identical. On the other hand, the two-day ratio $R_{1,2} = 2.3$ only captures one spread from the minimum low price and the maximum high price over days 1 and 2. In practical terms, $R_{1,2}$ is a unique log range, and even though it reflects the volatility over two days, the spread component in the range represents only one daily spread.}

Altogether, the two assumptions allow Corwin and Schultz (2012) to derive their estimator using simple algebraic treatment. We follow the same treatment below with slightly distinct notation. Suppose there is a spread $S\%$ which is evenly divided by sellers and buyers. Thus observed prices for buys are $0.5S\%$ higher than true values, and observed sells are $0.5S\%$ lower than the true value of a sell. Call $H^T_t$ the true daily high and $L^T_t$ as the true daily low. Naturally, true prices are not directly observed, since we do not observe $S\%$.

Consider the identity

$$R^O_t \equiv \ln \left( \frac{H^T_t (1 + 0.5S^{*})}{L^T_t (1 - 0.5S^{*})} \right)$$

where $S^{*}$ is the true (unobserved) per-cent spread. The observed log range differs from the true daily range when the spread is positive. As Parkinson (1980) has shown, assuming a geometric Brownian motion and a continuously traded asset (i.e., while the market is open there is continuous trade and no trading when markets are closed, in a sense that identification arises from price differentials), the moments of $R$ (precisely, w.r.t. the log range of actual values) can be calculated from\footnote{The asymptotic distribution of $R$ was first introduced by Feller (1951). Moreover, the quantities in (2.4) can be understood in terms of $E \left[ T^{-1} \sum_{i=1}^{T} R_i \right]$ and $E \left[ T^{-1} \sum_{i=1}^{T} R^2_i \right]$.}

$$E [R^p] = \frac{4}{\sqrt{\pi}} \left( 1 - \frac{4}{2^p} \right) \Gamma \left( \frac{p + 1}{2} \right) \left( 2\sigma^2_{HL} \right)^{p/2} \xi(p - 1) \quad (2.2)$$

and therefore

$$E [R] = \sigma_{HL} \sqrt{\frac{8}{\pi}}, \quad E [R^2] = \sigma^2_{HL} \ln 2 \quad (2.3)$$
are quantities of special interest. Call $S \equiv (1 + 0.5S^*)/(1 - 0.5S^*)$. After squaring both sides in (2.1), rearranging it and summing it over a consecutive-day interval, one obtains:

$$\sum_{t=1}^{2} (R_t^O)^2 = \sum_{t=1}^{2} (R_t^T)^2 + 2 \ln S \sum_{t=1}^{2} R_t^T + 2 \sum_{t=1}^{2} (\ln S)^2. \quad (2.4)$$

Note that the above is valid under the condition that the spread is constant in each consecutive two-day interval. After taking expectations,

$$\mathbb{E} \left[ \sum_{t=1}^{2} (R_t^O)^2 \right] = 2 \mathbb{E} \left[ \frac{1}{2} \sum_{t=1}^{2} (R_t^T)^2 \right] + 4 \mathbb{E} \left[ \frac{1}{2} \sum_{t=1}^{2} R_t^T \right] \ln S + 2 (\ln S)^2. \quad (2.5)$$

By inserting (2.3) into (2.5),

$$\mathbb{E} \left[ \sum_{t=1}^{2} (R_t^O)^2 \right] = 2 \sigma_{HL}^2 4 \ln 2 + 8 \sigma_{HL} \left( \frac{2}{\pi} \ln S + 2 (\ln S)^2 \right). \quad (2.6)$$

For simplicity, call $\beta \equiv \mathbb{E} \left[ \sum_{t=1}^{2} (R_t^O)^2 \right]$ and $\alpha' \equiv \ln((2 + S^*)/(2 - S^*)).$ The spread $S^*$ can be solved for such that the following relation holds:

$$S^* = \frac{2 \left( e^{\alpha'} - 1 \right)}{e^{\alpha'} + 1} \quad (2.7)$$

At this point, the only unknown random variables are $S^*$ (and thereby $\alpha'(S^*)$) and $\sigma$. Rearranging (2.6) in terms of $\alpha'$ and $\beta$ defines the expression for two consecutive days:

$$2 \sigma_{HL}^2 4 \ln 2 + 8 \sigma_{HL} \left( \frac{2}{\pi} \alpha' + 2 \alpha^2 - \beta \right) = 0. \quad (2.8)$$

After dividing (2.8) across by 2, adding $\sigma_{HL}^2 8/\pi$ to both sides and manipulating it such that

$$\left( \alpha' + 2 \sqrt{\frac{2}{\pi} \sigma_{HL}} \right)^2 = \sigma_{HL}^2 \left( \frac{8}{\pi} - 4 \ln 2 \right) + \frac{\beta}{2} \quad (2.9)$$

---

5) $\mathbb{E} [\ln S] = \ln S$ because the spread $S$ was assumed to be constant over a two-day period.

6) Corwin and Schultz (2012) use only $\alpha$ instead of $\alpha'$. I use the latter notation to distinguish between the still undetermined $\alpha'$ and the closed-form solution thereof later presented.
is defined, Corwin and Schultz (2012) consider only the positive root of (2.9), since the spread $S^*$ must be non-negative.\footnote{Note that $\alpha' \in (0, 1)$ if $S \in (0, 1)$.} Therefore

\[
\alpha' = \sqrt{\sigma_{HL}^2 \left( \frac{8}{\pi} - 4\ln 2 \right) + \frac{\beta}{2} - 2 \sqrt{\frac{2}{\pi} \sigma_{HL}}}. \quad (2.10)
\]

Likewise, we can manipulate $R_{t,t+1}$ in the same vein as with $R_t$. Thus

\[
\left( R_{t,t+1}^O \right)^2 = \left( R_{t,t+1}^T \right)^2 + 2R_{t,t+1}^T \ln S + (\ln S)^2 \quad (2.11)
\]
is the two-day interval counterpart of (2.4). By taking expectations of (2.11) and assuming that volatility increases proportionately in time, we have the expression for the two-day period:

\[
2\sigma_{HL}^2 4 \ln 2 + 8\alpha' \sqrt{\frac{\sigma_{HL}^2}{\pi}} + \alpha'^2 - \gamma = 0 \quad (2.12)
\]

where $\gamma \equiv \mathbb{E} \left[ \left( R_{t,t+1}^O \right)^2 \right]$. We can insert (2.10) into the above equation such that (2.12) becomes

\[
\sigma_{HL}^2 \left[ \frac{8}{\pi} \left( 2 - 2\sqrt{2} \right) + 4 \ln 2 \right] + 2\sigma_{HL} \sqrt{\frac{2}{\pi} \left( 2\sqrt{2} - 2 \right)} \sqrt{\frac{\sigma_{HL}^2}{\pi} \left( \frac{8}{\pi} - 4 \ln 2 \right) + \frac{\beta}{2} - \frac{\beta}{2} - \gamma} = 0. \quad (2.13)
\]

One can obtain closed-form solutions for $\alpha'$ and $\sigma$. It is still necessary to solve for $\alpha'$ since the final expression for the CSHL estimator is obviously (2.7) with a treatable solution $\alpha$ for $\alpha'$. The parameters $\beta$ and $\gamma$ can be estimated from data since they only involve observed prices. However, $\alpha'$ is still related to the unobservable $\sigma$. The solution to (2.10) and (2.13) can be simplified if the strict inequality below is ignored

\[
\mathbb{E} \left[ \frac{\sum_{t=1}^{T} R_{t}^2}{T} \right] \geq \left( \mathbb{E} \left[ \frac{\sum_{t=1}^{T} R_{t}}{T} \right] \right)^2 \quad (2.14)
\]

so that $\sigma_{HL}^2 4 \ln 2 \geq \frac{8}{\pi} \sigma_{HL}^2$ yields $4 \ln 2 \approx 8/\pi$. The suggestion to attain the lower bound in the Jensen’s inequality above is the last assumption put forward by Corwin and Schultz.
(2012) to derive their estimator. The expression (2.13) then becomes

$$\sigma^2_{HL} \left( \frac{8}{\pi} \left(3 - 2\sqrt{2}\right) \right) + 2\sigma_{HL} \sqrt{\frac{2}{\pi} \left(2\sqrt{2} - 2\right) \sqrt{\frac{\beta + \beta}{2} - \gamma}} = 0$$

(2.15)

hence

$$\sigma^2_{HL} + 2\sigma_{HL} B + C = 0$$

(2.16)

where $B \equiv \left(\sqrt{\beta} - \sqrt{\beta/2}\right) / \sqrt{8/\pi} \left(3 - 2\sqrt{2}\right)$ and $C \equiv (\beta/2 - \gamma)/\left[(8/\pi) \left(3 - 2\sqrt{2}\right)\right]$, can be solved in a similar manner as with (2.8)

$$\sigma_{HL} = \sqrt{B^2 - C - B}$$

(2.17)

which can be finally plugged into (2.10) to provide a closed solution to $\alpha'$:

$$\alpha = \sqrt{\frac{2\beta - \sqrt{\beta}}{3 - 2\sqrt{2}}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}}$$

(2.18)

for $\beta$ and $\gamma$ defined as before. The CSHL estimator is thus

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

(2.19)

which characterizes an easily-implementable estimator, without relying on estimation of further parameters other than the structural pair $\beta, \gamma$. In the next subsection, I summarize ad hoc empirical adjustments in Corwin and Schultz (2012) to accommodate some limitations assumed in the model just discussed.

2.2 Empirical shortcomings

Some of the theoretical assumptions sustained previously are unrealistic. For example, the expression (2.12) assumes that there are no overnight returns. Corwin and Schultz (2012) suggest (and implement) three adjustments to address estimation shortcomings considering a stock market framework. However, these modifications may not apply to futures trading. We discuss this possibility in sequence by giving an extended treatment to post-estimation adjustments in the context of CSHL.
Negative adjustment. The assumption that both the sum of two consecutive range ratios and a two-day ratio reflect twice the volatility of one day was explicitly modeled as

$$\mathbb{E}\left[ (R_{t,t+1}^T)^2 \right] = 2\mathbb{E}\left[ \frac{1}{2} \sum_{i=1}^{2} (R_{i}^T)^2 \right] = 2\text{Var}[R]. \quad (2.20)$$

If the observed two-day variance is sufficiently more than twice the variance in one day, the CSHL estimator might be negative. What sufficiently entails in this context will be discussed at a later point in this work. For now, the negative adjustment follows what is done in the literature for similar cases (such as with the Roll estimator), negative spreads estimates are set to zero. Particularly,

$$\text{CSHL} = \begin{cases} 
\frac{2(e^\alpha - 1)}{1 + e^\alpha}, & \text{if non-negative} \\
0, & \text{otherwise.}
\end{cases} \quad (2.21)$$

All else equal, a greater number of zero spreads leads to underestimates of average spreads.

Overnight returns adjustment. By construction, the two-day ratio $R_{t,t+1}$ incorporates the overnight return in $[t,t+1]$ as it corresponds to a joint open-to-close and close-to-close trading window. On the other hand, $R_t$ accounts only for open-to-close (intraday) returns. Although this follows from construction, it contradicts the uninterrupted trade assumption. Corwin and Schultz (2012) mention that stock prices fluctuate significantly over nontrading hours (though volatility is higher during trading hours), which is a well established fact in the finance literature. Accordingly, commodity futures prices also present significant overnight returns that need to be accounted for (Christoffersen et al. (2014)).

Nonetheless, available settlement prices pose an additional choice variable that one does not find necessary in equity markets. We choose settlement prices as de facto closing prices, since settlement, and not closing prices, are employed for mark-to-market margin payment, which affects market participants’ short-term cash flows.

The adjustment rule for overnight returns is exemplified in Figure (2.1). First, reconsider the daily closing price, $C_t$, defined previously. The settlement price, $S_t = g(C_t)$ is some

---

8Note that in the presence of significant overnight changes without any adjustment, CSHL estimates will be underestimated, potentially generating more negative estimates. Therefore, adjusting for overnight returns should decrease the number of negative estimates to be considered as zero.

9In addition to that, closing prices were not always available in futures prices data.
transformation $g$ of the closing price following a specific procedure for each commodity. Consider the prices collected at day 1 and the next-day observed prices. Since the low price at 2, $L_2$, is greater than the previous-day settlement $S_1$ by $\delta_1$, we adjust extrema prices on 2 by decreasing both $H_2$ and $L_2$ by $\delta_1$. The adjusted day 2 prices are now used as reference for day 3. In this case, the observed high $H_3$ is lower than the settlement at 2, $S_2$ by $\delta_2$, so that we adopt the converse procedure and increase both high and low prices by $\delta_2$. Note that settlement prices remain unchanged. Rather than assuming that the overnight return is simply the difference between settlement and next-day opening prices, we adjust more broadly for large price variations that might have been influenced by activity in nontrading hours.

**Liquidity adjustment.** Infrequent trading might underestimate daily highs and lows due to insufficient liquidity. Figure (2.2) illustrates the adjustments in all cases considered in Corwin and Schultz (2012). Objectively, lack of liquidity poses a concrete problem when $H_{t+1}^O = L_{t+1}^O = c$, where $c$ is some positive price. When that is the case, but $c \in [L_t^O, H_t^O]$, second-day trade prices are set to $H_{t+1}^O = H_t^O$ and $L_{t+1}^O = L_t^O$, rather than $c$. Furthermore, when there is no trade volume on $t$, the same procedure is employed. For when $c$ is outside the previous day range, i.e., $c - H_t^O = \varepsilon$ or $L_t^O - c = \varepsilon$, $\varepsilon > 0$, second-day extrema are adjusted w.r.t. $\varepsilon$. When the second-day price lies above the previous day high, the second-day pair is adjusted as $(H_{t+1}^O, L_{t+1}^O) = (H_{t+1}^O, L_{t+1}^O + \varepsilon)$. Conversely, when $L_t^O - c = \varepsilon$, the adjustment follows $(H_{t+1}^O, L_{t+1}^O) = (H_t^O - \varepsilon, L_{t+1}^O)$. 

---

\[^{10}\text{Price movements are largely influenced by liquidity.}\]
Part (a) illustrates when second-day extrema are identical and bounded by the previous day range. In this case, adjusted second-day high and low are set equal to the previous day values, $H^*_2 = H_1$ and $L^*_2 = L_1$. Part (b) represents second-day high and low above the previous day range by $\varepsilon$. Adjustment keeps $H_2$ and increase previous day low by $\varepsilon$. Similarly, part (c) adjusts for second-day extrema below previous range by scaling down previous day high by $\varepsilon$ and keeping the observed low.

Figure 2.2: Liquidity adjustment

The adjustment regarding null volume days is distinct from the following two alternatives aforementioned, for second-day same prices. When volume is equal to zero, not necessarily high and low values are identical. This means that replacing values in $t + 1$ with the range in $t$ when the correspondent volume on $t + 1$ is zero only increases the number of equal highs and lows if and only if $H_t^O = L_t^O$. In other words, one should first apply the procedure for volume and then observe if there were added daily equal extrema prices. Namely, that introduces identical extrema across two consecutive days, since the zero volume day replicates the previous range. On the other hand, extreme values can be identical in days with no trade volume. If the previous extrema are not equal, then the procedure reduces the number of equal prices.
Chapter 3

HISTORICAL SPREAD ESTIMATION

3.1 Data

The researcher in futures markets must always decide upon which rolling method to use in order to generate uninterrupted futures prices series. The importance of the rolling procedure choice shall not be overlooked, particularly given the role of liquidity provision in financial market microstructural features. In that sense, rolling contracts in a manner that captures liquidity as realistically as possible is crucial to properly reflect bid-ask spreads, since using a front contract that may not correspond to that market’s actual activity understates liquidity and potentially affect BAS estimates. We obtain daily prices data for six major commodity futures contracts: corn, soybeans, wheat, sugar, live cattle, and Brent crude oil covering the last half-century, as shown in Table (3.1). Continuous price series are constructed by rolling a front contract to the first deferred contract whenever the open interest in the latter becomes larger than in the former. This intends to capture actual market rolling movements, which precede the last trading day (another possible rolling method) and have been well documented (for example, see Irwin et al. (2011) and Wang et al. (2014), for agricultural markets).\footnote{An alternative, but identical in spirit, liquidity-based roll procedure would take into account volume rather than open interest.} Terminal points for all series are set to 08/31/2017.

3.2 Transaction costs in commodity markets

This section presents the analysis of historical spread estimates generated by the high-low estimator. Table (3.2) displays main aggregate results for all choices of ad-hoc empirical adjustments (liquidity, overnight returns, and negative estimates) meant to alleviate the restrictiveness of some assumptions embedded in the CSHL estimator. Liquidity adjust-
Table 3.1: Selected commodity futures contracts

<table>
<thead>
<tr>
<th></th>
<th>Obs.</th>
<th>Time horizon</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Corn</td>
<td>14,656</td>
</tr>
<tr>
<td>2</td>
<td>Soybeans</td>
<td>12,272</td>
</tr>
<tr>
<td>3</td>
<td>Wheat</td>
<td>14,649</td>
</tr>
<tr>
<td>4</td>
<td>Sugar</td>
<td>13,677</td>
</tr>
<tr>
<td>5</td>
<td>Live cattle</td>
<td>13,275</td>
</tr>
<tr>
<td>6</td>
<td>Brent oil</td>
<td>6,222</td>
</tr>
</tbody>
</table>

Notes: Data comes from the CME Group and Intercontinental Exchange (ICE). Rollover is chosen on open-interest switch basis.

...ments alter the dataset at the most fundamental level, as it replaces observed daily prices when trading activity is infrequent. Overnight returns adjustments (may) only modify daily ranges that will be mapped by $\beta$ and $\gamma$, and setting negative estimates to zero changes the relevant final sample for obtaining the average spread. The justification in Corwin and Schultz (2012) for implementing these adjustments is somewhat arbitrary.

The most sensitive adjustment, for liquidity, even yields internal inconsistencies in its implementation. To illustrate, imagine we observe four consecutive trading days, $t = 1, ..., 4$, and the following pairs of daily extrema $(H_t, L_t): (200, 191)_1, (195, 195)_2, (187, 180)_3$ (with no trading volume), and $(175, 175)_4$. The liquidity adjustment fits naturally: there are two days with identical daily prices and one day with no trading volume. But which one to deal with first? If we start by replacing the pair of prices on day 3 with the previous range, and then adjust for identical daily prices, the final, adjusted data will look like $(200, 191)_{t=1,2,3}$ and $(184, 175)_4$. On the other hand, if we first adjust identical prices according to the rules in Figure (2.2), and then replace the price at $t = 3$, the final data looks different: $(200, 191)_{1,2,3}$, and $(187, 180)_4$. Even if we were to impose a rule for substituting day 4 with the previous day range only if day 3 has positive volume, and then remove this rule before adjusting for same-day prices, it would take more steps (in an implementation sense) to achieve the original modified data. Further, if day 3 had identical prices, e.g. $(201, 201)_3$, should we treat it as a volume-zero adjustment first, or as an equal price case? The first option would make prices at $t = 3$ to be $(195, 195)_3$, though the second yields $(201, 192)_3$. This clearly shows the difficulty in adopting the integral liquidity adjustment in a fully consistent fashion.

The first part of Table (3.2) presents average high-low spread estimates without adjusting for overnight returns or liquidity. When averaging over the full sample, spreads are...
negative, as the frequency of negative estimates is always above 40%. Discarding negative observations prior to computing the average, produces positive, but greater estimates than setting negative spreads equal to zero. Hence, the full sample and positive-spreads-only constitute a lower and upper bound, respectively, for the estimates under the negativity adjustment. This pattern holds regardless of what combination of liquidity and overnight returns adjustments are chosen. In the last section of Table (3.2), all ad-hoc adjustments are jointly performed. As mentioned above, the liquidity adjustment might yield different modified prices depending on the order selected for its implementation. We first adjust for observations when trading volume is zero, then correct identical daily prices, whether they are within, above or below the previous-day range. Overall, different combinations of liquidity and overnight returns adjustments, for when negative spreads are set to zero, return similar estimates for each futures market. Results are robust when futures prices series are constructed with the last-trading day rolling procedure, as done for example by Marshall et al. (2011).
Table 3.2: Commodity futures high-low spread estimates

<table>
<thead>
<tr>
<th>Adjustments</th>
<th>No overnight</th>
<th>No liquidity</th>
<th>Overnight</th>
<th>No liquidity</th>
<th>Liquidity</th>
<th>No overnight</th>
<th>Overnight</th>
<th>Liquidity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.50</td>
<td>0.63</td>
<td>0.31</td>
<td>-0.04</td>
<td>0.63</td>
<td>0.35</td>
<td>-0.50</td>
<td>0.63</td>
</tr>
<tr>
<td>Max (%)</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
<td>4.94</td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>51</td>
<td>44</td>
<td>51</td>
<td>44</td>
<td>51</td>
<td>44</td>
<td>51</td>
<td>44</td>
</tr>
<tr>
<td>Soybeans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.50</td>
<td>0.68</td>
<td>0.35</td>
<td>-0.03</td>
<td>0.69</td>
<td>0.39</td>
<td>-0.49</td>
<td>0.68</td>
</tr>
<tr>
<td>Max (%)</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>48</td>
<td>41</td>
<td>48</td>
<td>41</td>
<td>48</td>
<td>41</td>
<td>48</td>
<td>41</td>
</tr>
<tr>
<td>Wheat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.47</td>
<td>0.74</td>
<td>0.38</td>
<td>-0.04</td>
<td>0.73</td>
<td>0.41</td>
<td>-0.47</td>
<td>0.74</td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>49</td>
<td>43</td>
<td>49</td>
<td>43</td>
<td>49</td>
<td>44</td>
<td>49</td>
<td>43</td>
</tr>
</tbody>
</table>
Table 3.2 (Continued): Commodity futures high-low spread estimates

<table>
<thead>
<tr>
<th>ADJUSTMENTS</th>
<th>No overnight</th>
<th>Overnight</th>
<th>Liquidity</th>
<th>Overnight</th>
<th>Liquidity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No liquidity</td>
<td>No liquidity</td>
<td>No liquidity</td>
<td>No liquidity</td>
<td>No liquidity</td>
</tr>
<tr>
<td>Sugar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.79</td>
<td>1.11</td>
<td>0.57</td>
<td>-0.06</td>
<td>1.12</td>
</tr>
<tr>
<td>Max (%)</td>
<td>10.30</td>
<td>10.30</td>
<td></td>
<td>10.30</td>
<td></td>
</tr>
<tr>
<td>Min (%)</td>
<td>-58.05</td>
<td>-14.04</td>
<td></td>
<td>-58.05</td>
<td></td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>49</td>
<td>42</td>
<td></td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Live Cattle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.38</td>
<td>0.50</td>
<td>0.25</td>
<td>-0.02</td>
<td>0.50</td>
</tr>
<tr>
<td>Max (%)</td>
<td>3.29</td>
<td>3.29</td>
<td></td>
<td>4.30</td>
<td></td>
</tr>
<tr>
<td>Min (%)</td>
<td>-25.45</td>
<td>-5.56</td>
<td></td>
<td>-25.45</td>
<td></td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>49</td>
<td>42</td>
<td></td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>Brent Oil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread (%)</td>
<td>-0.16</td>
<td>1.11</td>
<td>0.62</td>
<td>0.11</td>
<td>1.12</td>
</tr>
<tr>
<td>Max (%)</td>
<td>8.39</td>
<td>8.39</td>
<td></td>
<td>11.35</td>
<td></td>
</tr>
<tr>
<td>Min (%)</td>
<td>-28.84</td>
<td>-10.43</td>
<td></td>
<td>-28.84</td>
<td></td>
</tr>
<tr>
<td>Neg. (%)</td>
<td>44</td>
<td>40</td>
<td></td>
<td>43</td>
<td></td>
</tr>
</tbody>
</table>

Notes: Both table sections show estimated spreads obtained by the CSHL estimator from Corwin and Schultz (2012). Ad hoc adjustments follow the procedure described in the main text.
In general, spreads are of similar magnitude for grains, in the neighborhood of 0.35% (corn) and 0.42% (wheat). Sugar and Brent oil trading costs situate around 0.65% and spreads for live cattle futures are lower than grain markets, 0.29%. The frequency of negative estimates is relatively stable across commodities, with at least 40% of total 2-day CSHL estimates being negative. The direct effects of a high proportion of negative spreads towards the global behavior of the high-low estimator are unknown. Another interesting finding contained in Table (3.2) is the (absolute) size of minimum estimates compared to maximum spreads. For example, the minimum spread computed for corn, -65.14%, clearly shows an extremely anomalous behavior of the estimator.

How have trading costs behaved over time? Figure (3.1) presents historical spread estimates plotted as monthly averages. The absence of a clear downward pattern of trading costs across all commodity markets is puzzling. As a matter of fact, some historical series show even an increasing trend in trading costs, particularly in most recent decades. For corn, the median monthly average spread for the period 1996-2016 is 0.47%, although the average monthly median between 1959 and 1995 is 0.23%. The trends in other markets emerge less clearly, with a fairly stable behavior in soybeans and live cattle from the late-1970s onwards, and nonetheless significant oscillation in all series.

The overwhelming understanding in the literature concerning trading costs in financial markets, including commodity futures, is that migration to electronic trading leads to much lower transaction costs (to mention few studies, Pagano and Roell (1996), Tse and Zabotina (2001), Martinez et al. (2011), Pirrong (1996), Frank and Garcia (2011), Wang et al. (2014), Shah and Brorsen (2011), and Menkveld (2016)). Our results, in lieu with previous findings, leads to a questioning of the reliability of the CSHL estimator. The assessment of potential sources of measurement error which might be contaminating historical spread trends is extremely limited. One may conjecture that the CSHL estimator may be capturing volatility over time, which could partially explain the overall nondecreasing trend seen in Figure (3.1). In the next section, we investigate the small sample properties of the CSHL estimator, identifying sources of data variation that may cause the estimator to behave poorly.
Figure 3.1: Commodity futures historical CSHL spread estimates
Chapter 4

MEASUREMENT ERROR IN THE CSHL ESTIMATOR

4.1 The importance of the volatility-to-spread ratio

In this subsection, a modification of the simulation performed by Corwin and Schultz (2012) (CS) is conducted to identify possible sources of measurement error in the high-low spread estimator. Even under ideal conditions, i.e. under a price generating process that nearly replicates the CSHL model assumptions, there is reason to doubt the accuracy of computed sample means of spreads.\(^1\) Since standard asymptotic analysis would depend on the undefined relationships between the high-low estimator and theoretical quantities, such as the true spread, simulations are useful as a first approach. Throughout the analysis, let \(\bar{\text{CSHL}} \equiv N^{-1} \sum_{i=1}^{N} \hat{\text{CSHL}}_{i}\), where \(\hat{\text{CSHL}}_{i}\) represent CSHL estimates obtained from (2.19) and adjusted for negative values. Measurement error is defined as the simple distance \(\bar{\text{CSHL}} - S^{*}\), \(S^{*}\) the true spread, a constant and known choice variable. Ideally, no systematic error should arise from changes in certain parameters that regulate how prices, and therefore true bid-ask spreads, behave. That is, confidence in the use of the CSHL estimator depends on its robustness to a range of true spreads and volatility regimes that correspond to observed market realities.

I expand the ideal environment simulated in CS by replicating it and allowing volatility to vary. The interest in understanding small sample effects of volatility stems from the manner the CSHL estimator captures liquidity. By assuming a constant spread for each two-day interval, the high-low estimator captures transitory volatility, hence incorporating liquidity more broadly than simply the bid-ask spread. What first appears to be a useful property, capturing additional components of liquidity costs might decrease the relative information relevance of true spreads in determining the CSHL. Following Corwin and Schultz (2012),

\(^1\)This concern is mainly motivated by Section IV in Corwin and Schultz (2012), where estimated spreads (after setting 2-day negative estimates to zero) seem to be significantly larger for small true spreads than for high spread levels. There is also some evidence suggesting that volatility impacts the degree of accuracy of high-low spreads, yet its effect remains largely unexplored.
define the efficient price $p$ from the following DGP for efficient returns:

$$\tilde{\sigma} x = \log \left( \frac{p_m}{p_{m-1}} \right) \tag{4.1}$$

where $p_m$ represents a given price at minute $m$, $m = 1, ..., M$, $x \sim \mathcal{N}(0, 1)$ and $\sigma$ defined as one of the experimental choice variables ($\tilde{\sigma} \equiv \sigma / \sqrt{M}$). The general dimensions of the simulation determine the random vector $x = (x_1, ..., x_K)$, where $K = (390 = M \times 21) \times N$ represents the total temporal dimension: $N$ months with 21 days composed of 390 minutes each. Every random variable $x$ is matched with a correspondent value in a $(390 \times 21) \times N$ matrix $P$, where $p_{1,j} = 100 \exp(\sigma x_{1,j})$, $j = N$, and $p_{i \neq 1} = p_{i-1} \exp(\sigma x_{i,j})$. That is identical to set the price before each first minute of each month’s first day to 100 and for all $M$ prices to follow (4.1).

After generating intraday prices for $N$ months, we need to obtain daily extrema to compute parameters $\beta$ and $\gamma$ and in turn the CSHL estimator. Given the assumption that daily high (low) prices are buyer- (seller-) initiated trades, the ask (bid) quotes per minute calculated as below will provide high and low prices for each consecutive interval of length 390, for all $21 \times N$ days:

$$\text{Bid}_m = p_m (1 - S\%/2) \tag{4.2}$$

$$\text{Ask}_m = p_m (1 + S\%/2) \tag{4.3}$$

where the true spread $S \equiv S^*$ is defined ex ante.

Replication results closely meet CS’s third column in Table I, when $\sigma = 0.03$. Small differences might be regarded to the selection criteria of $x$’s drawn from the standard normal used in the DGP. The variable component of the temporal dimension, $N$ (number of months), should be appropriate in order for estimates to be reliable. Rather than simulating 10,000 months as in Corwin and Schultz (2012), $N = 100$ months are enough to ensure stable estimates. Negative spreads are set to zero at daily frequency and high-low estimates are averaged over the entire time horizon, CSHL. The environment simulated herein enables the CSHL estimator to perform in optimal fashion. Table (4.1) presents sample means and frequency of negative and overestimated high-low spreads for each combination of $\langle S, \sigma \rangle$.

Clearly, for high true spreads (e.g. 8%), volatility plays a modest role in introducing measurement error in estimates. However, when true spreads are small - under 1% - variability in estimates is significant. Particularly, under a combination of high volatility and a
small true spread, or a high volatility-to-spread ratio, CSHL estimates overstate the actual spread as much as 16-fold, yielding far from “ideal” results. The immediate conclusion would relate the increasing poor accuracy in average CSHL estimates as we move towards the right lower portion of Table (4.1) to the joint effect of \((S, \sigma)\). Nonetheless, this relationship is yet to be established. The direction of increasing positive measurement error points to both smaller true spreads and higher volatility values, making it hard to disentangle the effects of overestimated spreads occurring at a higher frequency, or insufficient (in number or magnitude) viable below-\(S\) estimates in order to offset above-\(S\) estimates. The additional statistics in Table (4.1) aim to shed some light upon the individual contribution of both possibilities. The values corresponding to the frequency of overestimation, reported in Table (4.1), fail to support variations in overstated spreads’ frequency as a plausible rationale.\(^2\) For example, CSHL estimates that return an average value of 3.29% when \(S = 0.2\%\) and \(\sigma = 8\%\), exceed the true spread in 57.5% of estimated values, whereas the average high-low spread of 3.23% (\(S = \sigma = 3\%\)) arises from 56.07% overstated estimates. If more frequent overestimated spreads is an unsatisfactory explanation for the nature of the measurement error, alternatively, the empirical distribution of unadjusted high-low estimates might indicate whether viable below-\(S\) spreads are insufficient to offset overestimated spreads.

In Figure (4.1), we compare the distribution of CSHL estimates when daily volatility is chosen as \(\sigma = 8\%\), but true spreads are 8% or 0.2%.\(^3\) A first striking distinction between sub-panels (a) and (c) is the number of below-\(S\), but positive (viable), estimates when \(S = 8\%\) (represented by blue dots), compared to the same positive understated spreads when \(S = 0.2\%\) (blue dots in (c) are virtually indistinguishable from negative spreads). The straightforward consequence arising from this discrepancy is fewer observations “from below” counting towards the true-mean convergence. Namely, even though the density of overstated estimates in (b) is similar to (d), there is a significant mass of below-\(S\) positive estimates in (a) (blue dots) offsetting large positive outliers, which contributes to an overall distribution centered relatively close to the true mean, \(\overline{\text{CSHL}} = 0.861\) for \(S = 0.8\). On the other hand, in (c), large negative outliers are reduced (in absolute terms) when adjusted to zero in one of the post-estimation procedures, and without enough mass of estimates between 0 and 0.002, large positive outliers heavily influence the high-low estimator. The dynamics reported in these two settings is commonplace across different levels of distinct spreads, for all volatility levels.

\(^2\)The frequency of overstated spreads, \(\overline{\text{CSHL}} > S\), is computed for CSHL estimates corrected for negative values, i.e., set to zero. Positive measurement error of this form with negative spreads set to zero yields no difference from raw spreads being used instead.

\(^3\)The analysis holds for all volatility levels, although for \(\sigma = 8\%\) the results discussed above are more prominent.
This suggests that overestimation increasing in lower true spreads and greater volatility as seen in Table (4.1) could be explained by a lack of positive underestimated estimates to counterweight large positive spreads (viable below-S estimates). Greater volatility decreases accuracy, but precision losses become more acute when true spreads are low and a greater number of estimates turn out negative. More concisely, measurement error in the high-low estimator increases in the volatility-to-spread ratio.

<table>
<thead>
<tr>
<th>Spread (S)</th>
<th>0.5%</th>
<th>1%</th>
<th>3%</th>
<th>5%</th>
<th>8%</th>
</tr>
</thead>
<tbody>
<tr>
<td>8%</td>
<td>7.84%</td>
<td>7.77%</td>
<td>7.75%</td>
<td>7.94%</td>
<td>8.61%</td>
</tr>
<tr>
<td></td>
<td>(0.57%)</td>
<td>(1.67%)</td>
<td>(3.86%)</td>
<td>(8.96%)</td>
<td>(17.91%)</td>
</tr>
<tr>
<td></td>
<td>[53.31%]</td>
<td>[53.69%]</td>
<td>[54.84%]</td>
<td>[55.50%]</td>
<td>[56.07%]</td>
</tr>
<tr>
<td>5%</td>
<td>4.87%</td>
<td>4.84%</td>
<td>4.95%</td>
<td>5.39%</td>
<td>6.33%</td>
</tr>
<tr>
<td></td>
<td>(1.19%)</td>
<td>(2.33%)</td>
<td>(8.43%)</td>
<td>(17.96%)</td>
<td>(25.58%)</td>
</tr>
<tr>
<td></td>
<td>[53.50%]</td>
<td>[53.98%]</td>
<td>[55.36%]</td>
<td>[56.07%]</td>
<td>[56.26%]</td>
</tr>
<tr>
<td>3%</td>
<td>2.90%</td>
<td>2.90%</td>
<td>3.23%</td>
<td>3.87%</td>
<td>4.97%</td>
</tr>
<tr>
<td></td>
<td>(2.14%)</td>
<td>(3.33%)</td>
<td>(17.96%)</td>
<td>(26.25%)</td>
<td>(32.16%)</td>
</tr>
<tr>
<td></td>
<td>[53.74%]</td>
<td>[54.74%]</td>
<td>[56.07%]</td>
<td>[56.26%]</td>
<td>[56.79%]</td>
</tr>
<tr>
<td>1%</td>
<td>0.98%</td>
<td>1.08%</td>
<td>1.78%</td>
<td>2.56%</td>
<td>3.74%</td>
</tr>
<tr>
<td></td>
<td>(6.05%)</td>
<td>(17.96%)</td>
<td>(33.11%)</td>
<td>(36.73%)</td>
<td>(39.16%)</td>
</tr>
<tr>
<td></td>
<td>[55.31%]</td>
<td>[56.07%]</td>
<td>[56.88%]</td>
<td>[57.12%]</td>
<td>[57.27%]</td>
</tr>
<tr>
<td>0.5%</td>
<td>0.54%</td>
<td>0.70%</td>
<td>1.47%</td>
<td>2.26%</td>
<td>3.45%</td>
</tr>
<tr>
<td></td>
<td>(17.96%)</td>
<td>(26.58%)</td>
<td>(37.92%)</td>
<td>(39.59%)</td>
<td>(40.78%)</td>
</tr>
<tr>
<td></td>
<td>[56.07%]</td>
<td>[56.60%]</td>
<td>[57.22%]</td>
<td>[57.36%]</td>
<td>[57.41%]</td>
</tr>
<tr>
<td>0.2%</td>
<td>0.32%</td>
<td>0.51%</td>
<td>1.30%</td>
<td>2.10%</td>
<td>3.29%</td>
</tr>
<tr>
<td></td>
<td>(31.63%)</td>
<td>(36.73%)</td>
<td>(40.54%)</td>
<td>(41.45%)</td>
<td>(41.78%)</td>
</tr>
<tr>
<td></td>
<td>[56.79%]</td>
<td>[57.12%]</td>
<td>[57.41%]</td>
<td>[57.46%]</td>
<td>[57.50%]</td>
</tr>
</tbody>
</table>

**Notes:** Estimated spreads are average values. Values in (∗) represent frequency of negative daily estimates. Values in [∗] correspond to the frequency of overestimation (CSHL > S) for each combination of volatility and true spread.

Insofar as simulated high-low estimates revealed that high volatility and small true spread combinations could lead to potentially large measurement error in applied work, such as the likely misleading historical estimates in Figure (3.1), the usefulness of such findings is hampered by the reliance on a priori knowledge of the true spread S∗. As (theoretical) volatility itself is insufficient to fully account for performance issues, the dependence on some empirically observed phenomenon that is highly correlated to a small true spread could constitute a convenient venue. With that respect, the frequency of negative estimates in
Subfigure (a) displays CSHL unadjusted spreads obtained in the Monte Carlo experiment under choice variables $S = \sigma = 8\%$. The values in gray are overestimated spreads, $\hat{\text{CSHL}}_i > 0.08$, where $i/N = 56.07\%$, $N = 2,099$. Values depicted in blue correspond to underestimated positive spreads, $0.08 < \hat{\text{CSHL}}_i < 0$ and the few points below zero represent negative spreads (adjusted to zero in order to obtain average values as in Table (4.1)). Negative spreads account for 17.91\% of total estimates. In subfigure (b), kernel density estimates generate the distribution of CSHL spreads, with the true value $S = 0.08$. We use a Gaussian kernel density estimator to select a bandwidth of 0.01606. Subfigure (c) displays CSHL unadjusted spreads obtained in the Monte Carlo experiment under choice variables $S = 0.2\%$ and $\sigma = 8\%$. The values in gray are overestimated spreads, $\hat{\text{CSHL}}_i > 0.002$, where $i/N = 57.5\%$, $N = 2,099$. The few points depicted in blue correspond to underestimated positive spreads, $0.002 < \hat{\text{CSHL}}_i < 0$ and values below zero again represent negative spreads. Negative spreads account for 41.78\% of total estimates. In sub-figure (d), kernel density estimates generate the distribution of CSHL spreads, with the true value $S = 0.002$. A Gaussian kernel density estimator selected a bandwidth of 0.01574.

Figure 4.1: CSHL estimates under almost ideal conditions
Table (4.1) appears to closely follow the direction of measurement error. Positively influenced both by higher volatility and smaller true spreads, the proportion of high-low estimates which turn out negative increases from 0.57% when $S = 8\%$ and $\sigma = 0.5\%$, to almost 42% for $S = 0.2\%$ and $\sigma = 8\%$. Altogether, the occurrence of negative spreads might indicate the presence of effects of a small true spread and high volatility in the CSHL estimator. While the role of negative high-low estimates and how to deal with them have played some role in the literature, there is no clear mechanism that explains why negative spreads arise, as the simple diagnosis of positive covariance of returns does for the Roll (1984) covariance spread estimator.\footnote{An example can be found in Shane Corwin’s reply to Kim and Lee (2014): https://www3.nd.edu/ scorwin/documents/DealingwithNegativeValues.pdf.} In the next subsection, we develop a formal treatment of which circumstances lead to high-low spreads in the negative region. The usefulness of determining such cases lies in potentially using the frequency of negative spreads to indirectly assess measurement error.

### 4.2 The behavior of $\alpha$, $\beta$ and $\gamma$

The understanding of negative high-low spread estimates remains limited to the original work of Corwin and Schultz (2012), where it is argued that if variance proportionality (i.e. linear in time) fails to hold in practice, estimates may be negative. As a consequence, they document the frequency of negative spreads and advise setting daily negative estimates to zero prior to obtaining CSHL as the most preferable empirical adjustment to deal with negative high-low spreads. In order to determine under which conditions (2.19) is negatively determined, we must verify the behavior of its parameters $\alpha$, $\beta$ and $\gamma$. The CSHL estimator as in (2.19) can be written equivalently as

$$S \equiv 2 \tanh \left( \frac{\alpha}{2} \right)$$  \hspace{1cm} (4.4)

where $\tanh (\cdot)$ is the hyperbolic tangent function. The hyperbolic function representation is useful since the range of $S$ follows straightforwardly: $S \in (-2, 2)$, for which only $[0, \bar{S})$, for some $\bar{S} > 0$, would be admissible, although negative estimates are empirically recurrent. Indeed, as Figure (4.1) shows, there is a significant magnitude asymmetry between positive and negative outliers, suggesting that negativity not only occurs frequently, but the underlying causes of it seem to be strong enough to give degenerated spreads with lack of sensibility to marginal improvements in market conditions. Since the spread is only determined by $\alpha$, our
problem simplifies to characterize the array of cases for which $\alpha$ is negative. The expression in (2.18) can be simplified so that the relationship between $\alpha$ and its parameters $\beta$ and $\gamma$ is unequivocal:

$$\alpha (\beta, \gamma) = \left( \sqrt{\beta} - \sqrt{\gamma} \right) \Theta,$$

(2.18′)

where $\Theta = 1 + \sqrt{2}$. The simplified parameter $\alpha (\cdot)$ imposes a straightforward negativity condition: $\beta < \gamma$. Hence, determining all cases for which $\gamma$ exceeds $\beta$ suffices to determine which empirical estimates of CSHL turn out negative. The parameters $\beta$ and $\gamma$ can be more fundamentally understood as functions whose arguments are uniquely determined by observed data. Let $\mathcal{G}_t$, $t = 1, 2$, be partitioned into $\mathcal{H}_t$ and $\mathcal{L}_t$ such that for all $h_t \in \mathcal{H}_1$ and $l_t \in \mathcal{L}_1$, $h_t > l_t$. Now, let $\beta : \mathcal{H}_1 \times \mathcal{H}_2 \times \mathcal{L}_1 \times \mathcal{L}_2 \rightarrow \mathbb{R}_+$,

$$\beta \left( H^O_1, L^O_1, H^O_2, L^O_2 \right) = \sum_{t=1}^{2} \left( \ln \left( \frac{H^O_t}{L^O_t} \right) \right)^2$$

(4.5)

where $\mathcal{L}_1 \subset \mathbb{R}_+$ is bounded from above by some $H^O_1 \in \mathcal{H}_1$, and similarly $\mathcal{H}_2 \subset \mathbb{R}_+$ is bounded from below by some $L^O_2 \in \mathcal{L}_2$. The function $\gamma \left( H^O_1, L^O_1, H^O_2, L^O_2 \right)$ has the same domain characterization as $\beta (\cdot)$, following the functional form previously defined:

$$\gamma \left( H^O_1, L^O_1, H^O_2, L^O_2 \right) = \left( \ln \left( \frac{\max_{t=1,2} \{ H^O_t \}}{\min_{t=1,2} \{ L^O_t \}} \right) \right)^2.$$

(4.6)

For simplicity of exposition, let us omit the superscript for observed prices hereafter. The negativity condition (NC) of the high-low estimator is given by

$$\left( \ln \left( \frac{\max \{ H_1, H_2 \}}{\min \{ L_1, L_2 \}} \right) \right)^2 > R^2_1 + R^2_2.$$

(4.7)

The inequality in (4.7) would be a simple relationship to establish, were $\gamma$ not indirectly determined by all daily extrema, which in turn vary with the daily ranges. Sufficiently large changes in one argument, viz. increasing the second-day high by 20%, could change the $\gamma$

---

5 Please, see the Appendix for detailed analyses of results in this subsection.

6 I use the above notation to stress the importance of understanding that variations in one-day high, for example, might not affect the same day low price. In other words, the function $\beta (\cdot)$ chooses a first-day low value $L^O_1$ from $\mathcal{L}_1$, up to the point where $L^O_1$ cannot be greater than $H^O_1$, the same day high. Further, changes in $L^O_1$ over $(0, H^O_1)$, are autonomous from changes of $H^O_1$, as the latter only modifies the upper bound of $L^O_1$. 

27
function that selects \( \langle H_1, L_2 \rangle \) to \( \langle H_2, L_2 \rangle \). Consequently, we must consider variations in daily ranges that guarantee the existence of a \( \bar{\gamma} \) for which NC is always stable,

\[
\bar{\gamma}(H_1, H_2, L_1, L_2) > (R_1(H_1, L_1))^2 + (R_2(H_2, L_2))^2,
\]

(4.8)
since \( \gamma \) still maps the same daily high and low variables onto the positive real line. The condition above in addition to the stability restriction allow us to fully describe the possible solutions for the negativity of \( \alpha \). As our ultimate goal with this analysis is to provide a better understanding of the behavior of the high-low estimator in real world markets, a common feature of financial prices data represents an important complicating aspect for the number of cases to be considered in determining (4.8): the occurrence of consecutive-day identical prices, including daily high and low values. Note that the number of observed choices of the function \( \gamma \),

\[
\gamma(H, L) = \begin{cases} 
\gamma(H_1, L_1), & \text{if } H_1 \geq H_2 \text{ and } L_1 \leq L_2 \\
\gamma(H_2, L_1), & \text{if } H_2 \geq H_1 \text{ and } L_1 \leq L_2 \\
\gamma(H_1, L_2), & \text{if } H_1 \geq H_2 \text{ and } L_2 \leq L_1 \\
\gamma(H_2, L_2), & \text{if } H_2 \geq H_1 \text{ and } L_2 \leq L_1 
\end{cases}
\]

(4.9)
are indeed determined over nine different two-day-relative extrema combinations. As Figure (4.2) shows, we may define, for analytical clarity, an arbitrary “major” case that determines \( \gamma \) in terms of solely strict inequalities, and for each one of such cases let there exist other three possible “minor” situations closely related of combinations of high and low prices that determine the same observed parameters chosen by \( \gamma \). For example, suppose the high prices in two consecutive days are identical, but the first-day low is smaller than the second-day low price. In this case, both \( \gamma'(H_1, L_1) \) and \( \gamma''(H_2, L_1) \) are equivalent \( \gamma \)-representations of the data, as simply observing \( \gamma' \) or \( \gamma'' \) fails to reveal the relationship between \( H_1 \) and \( H_2 \) (\( H_1 \geq H_2 \), or \( H_2 \geq H_1 \), or both). Let major cases be unique for each distinct configuration of \( \gamma \), and indexed by \( i = 1, ..., 4 \). Minor cases represent 5 possibilities which can determine multiple \( \gamma \) and are indexed by \( j = A, B, ..., E \).\(^7\) We proceed by analyzing the appropriateness of NC for each major case. Minor cases can be seen as limiting special situations and are fairly simple to verify for the negativity condition. Thus the proofs for Cases A-E are contained in the Appendix.

\(^7\)The combinations of daily high and low prices are: (A) \( H_1 = H_2, L_1 < L_2 \); (B) \( H_1 = H_2, L_2 < L_1 \); (C) \( L_1 = L_2, H_1 > H_2 \); (D) \( L_1 = L_2, H_2 > H_1 \); and (E) \( H_1 = H_2, L_1 = L_2 \).
Assume that daily high and low prices are well defined, \( H_t > L_t > 0, \ t = 1, \ldots, N \), so that \( \beta \) is always greater than zero.

**CASE 1. Second-day high and low are contained in first-day extrema: \( H_1 > H_2 \) and \( L_1 < L_2 \). In this case, the CSHL estimator is never non-positive.**

Say there is a \( \tilde{\gamma}_1 \) for which

\[
\tilde{\gamma}_1 = \left( \ln \left( \frac{H_1}{L_1} \right) \right)^2
\]

holds. Naturally, from the definition of \( \gamma (\cdot) \), \( \tilde{H}_1 > H_2 \) and \( \tilde{L}_1 < L_2 \) follows. Let \( L_2 \in (\tilde{L}_1, H_2) \) and \( H_2 \in (L_2, \tilde{H}_1) \). Therefore, the negativity condition yields:

\[
\left( \ln \left( \frac{H_1}{L_1} \right) \right)^2 > \left( \ln \left( \frac{H_1}{L_1} \right) \right)^2 + \left( \ln \left( \frac{H_2}{L_2} \right) \right)^2
\]

which is clearly inadmissible, since \( H_i > L_i \), for \( i = \{1, 2\} \).

**CASE 2. Second-day high and low are greater than first-day extrema: \( H_1 < H_2 \) and \( L_1 < L_2 \). In this case, the CSHL estimator is more likely to be negative as \( H_2 \) becomes larger and \( L_2 \to H_1 \), if \( L_2 \in (L_1, H_1) \). When \( L_2 \in (H_1, H_2) \), the spread is always negative.**

We first consider when the second-day low is greater than the first-day high (that is, no daily range is contained in any portion of the other range). Suppose \( \tilde{L}_2 > \tilde{H}_1 > \tilde{L}_1 \) and \( \tilde{H}_2 \in (\tilde{H}_1, \tilde{H}) \), such that the negativity condition (NC) holds with \( \tilde{\gamma}_2 \):

\[
\tilde{\gamma}_2 = \left( \ln \left( \frac{H_2}{L_1} \right) \right)^2
\]

In this case, the NC translates as:

\[
\left( \ln \left( \frac{H_2}{L_1} \right) \right)^2 > \left( \ln \left( \frac{H_1}{L_1} \right) \right)^2 + \left( \ln \left( \frac{H_2}{L_2} \right) \right)^2
\]

29
Note that, because $L_1 < L_2$, we can write $L_1$ as $L_2 \equiv \psi L_1$, where $\psi > 1$. Likewise, $H_2 \equiv \delta H_1$, $\delta > 1$. Accordingly, (4.13) becomes

$$(\ln \psi + R_1)^2 > R_1^2 + \left( \ln \left( \frac{\delta}{\psi} \right) + R_1 \right)^2.$$  

(4.14)

The expression above can be reduced to

$$\ln \delta \ln \psi > \frac{1}{2} \left( \ln \left( \frac{H_1}{L_2} \right) \right)^2.$$  

(4.15)

Now we use our last assumption. Since $L_2 > H_1$, $L_2 \equiv \omega H_1$ follows, with $\omega < \delta$, as a result of $H_2 > L_2$. Thus, (4.15) yields

$$\frac{2 \ln \delta \ln \omega + 2 \ln \delta R_1}{(\ln \omega)^2} > 1$$  

(4.16)

for which $\ln \delta \ln \omega > (\ln \omega)^2$ always holds, because $\delta > \omega$ and the functions are strictly increasing in their arguments (as sufficiency). Let us move to the possibility of $L_2 \in (L_1, H_1)$.

Suppose we decide to increase $H_2$ above $H_1$ until some upper bound $\overline{H}$, i.e., $\overline{H}_2 \in (\overline{H}_1, \overline{H}]$, while maintaining $L_2$ fixed at $\overline{L}_2 > \overline{L}_1$, such that there exists a $\overline{\gamma}_2$ at which the inequality in (4.15) determines when $\overline{\gamma}_2 > \beta$.  

30
Figure 4.2: General cases of combinations with daily high and low prices.

Observed $\gamma$

$$\left(\frac{H_1}{L_1}\right)^2$$

$$\left(\ln\frac{H_1}{L_1}\right)^2$$

$$\left(\frac{H_2}{L_2}\right)^2$$

$$\left(\ln\frac{H_2}{L_2}\right)^2$$
The condition in (4.15) states that the joint vertical distance between $H_1, H_2$ and $L_1, L_2$ must compensate for the relative distance between the daily prices precluded by $\gamma$ in order to sustain NC. One may still require a more clear condition to observe negative spreads under the assumptions of Case 2. The clumsy relationship can be better understood if we use a numerical example. Define a function $G(H_2, L_2; 100, 90) > 1$ for whenever the negativity condition in Case 2 holds with $L_2 \in (L_1, H_1)$. This includes, for example, prices $(100, 90)_1$. Now, having established a value $\overline{H}_2$ at which NC is stable, how close to $H_1$ can we decrease $H_2$ and how much can we move it away from $H_1$ so that CSHL is always negative? How about the second-day low? We show below that even if the daily ranges are identical $(R_1 = R_2)$, if they are sufficiently far apart vertically, the spread will be negative. By increasing $H_2$ away from a $H^*_2$ at which NC holds with equality, or by increasing $L_2$ on $(L_1, H_1)$, the negativity condition is more easily attained. Likewise, as we showed before, by making both daily ranges disjoint, i.e., $L_2 > H_1$, the spread is always negative, even for $H_2$ arbitrarily close to $L_1$.

In Figure (4.3), we set the first day high and low as 100 and 90, respectively. The horizontal line $G = 1$ defines the portion of $G(H_2; L_2, 100, 90)$ that ensures $\gamma > \beta$. Some features are noteworthy mentioning. When the second-day low is $L_2 = 99$, the same-day high that attains $G$ is given by $H^*_2 \approx 100.05$. Hence, even though the second-day maximum is only 0.05% higher than $H_1$, a second-day low that is 10% higher than $L_1$ establishes the negativity condition for all $H_1$ slightly larger than 100.05. By the same token, the first-day (log) range is $\ln(100/90) = 0.105$ and the second-day range is $\ln(100.05/99) = 0.010$, ten times lower than $R_1$, yet the relation between both is sufficient to rule out positive spreads in the context presented. This shows that even slight price changes across two days can be enough to result in negative spreads.

**Case 3.** Second-day high and low are less than first-day extrema: $H_1 > H_2$ and $L_1 > L_2$. In this case, the CSHL estimator is more likely to be negative as $L_2$ becomes smaller and $H_2 \to L_1$, if $H_2 \in (L_1, H_1)$. When $H_2 \in (L_2, L_1)$, the spread is always negative.

This is symmetric to Case 2. The negativity condition is:

$$\ln \delta^{-1} \ln \psi^{-1} > \frac{1}{2} \left( \ln \left( \frac{H_2}{L_1} \right) \right)^2$$

(4.17)
Figure 4.3: Numerical example of Case 2

The figure shows the behavior of $G$, defined as 

$$G = \frac{2 \ln (H_2/100) \ln (L_2/90)}{(\ln (100/L_2))^2},$$

evaluated at $H_1 = 100$ and $L_1 = 90$, in the domain $(H_1, 100.5]$ for $H_2$. $G$ increases monotonically ($\partial G(\cdot)/\partial H > 0$) as we move the second-day high further away from $H_1 + \varepsilon$, $\varepsilon$ very small. Case 2 determines negative spreads whenever $G$ exceeds the support $G = 1$, so that the negativity condition (NC) holds for all $H_2 > H_2^\ast$, when $L_2 = 99$ and for all $H_2 > H_2^{\ast\ast}$, when $L_2 = 98$. As $L_2 \to H_1$, the speed of convergence of the high low spread estimate to 0 increases; the effect being contrary for smaller values of $L_2$.

**Case 4. Second-day high and low contain first-day extrema: $H_1 < H_2$ and $L_1 > L_2$. In this case, the CSHL estimator is never non-positive.**

Lastly, Case 4 is symmetric to Case 1.

In the simulations performed in the previous subsection, the frequency and positioning of each main case is identical for all different choices of volatility and true spread levels. Cases 2 and 3, which may induce negative spreads, correspond to 41% and 39% of all situations for consecutive-day prices, respectively. Cases 1 and 4 each occur in 10% of the sample. This is a direct consequence of the underlying process that selects the random variable present in the data generating process. Increased levels of volatility widen the vertical distance of daily extrema prices, but the relative two-day position remains unchanged. For example, on the sixth simulated day, the high and low prices generated under $S = 8\%$ and $\sigma = 3\%$, constituted the pair (89, 102), and the following daily-pair, (93, 105). The same couple of pairs under a DGP of $S = 8\%$ and $\sigma = 8\%$ returns (79, 100) on day 6 and (88, 107).
on day 7. On both cases, the second-day high is greater than the first-day maximum, and
the second-day low is also greater than the first-day minimum. That corresponds to Case
2. The practical difference lies in the negative spread induced by the occasion with higher
volatility, yet the first occurrence returned a positive spread estimate. By extending the
reasoning to all cases 2 and 3 across distinct spread-to-volatility ratios, differences in the
frequency of negative estimates are explained by changes in two-day prices corresponding to
cases 2 and 3 that become sufficient to uphold the negativity condition.

According to the framework presented, the effect of the true spread size in distorting
the CSHL measure via negative estimates is simply increasing the vertical distance in each
day’s range equally (if one assumes the each daily extreme is grossed up by half of the
spread). As the spread is assumed constant over each consecutive two-day interval, adding
the spread to our analysis does not impact the determination of NC. This suggests that
the much stronger measurement error found in estimates attempting to proxy small spread
levels may not be generated by the same mechanism that determines negative spreads.
In such case, volatility impacts how far CSHL estimates situate from the true spread,
and small actual spreads make more understated spreads become negative for all volatility
levels. This ultimately reveals a low precision of high-low estimates, for all spread levels.

In the next subsection, we quantify measurement errors arising in CSHL estimation
empirically. With TAQ data for corn futures, it is possible to obtain small true spreads
(using the effective spread as the high-frequency benchmark) and variable volatility over the
years to assess whether simulation predictions also hold in the data.

4.3 Actual trading costs and high-low spread estimates

In this subsection, we test the CSHL estimator performance against high frequency data
which allows us to observe actual trading costs. We are particularly interested in analyzing
the behavior of the high-low spread estimator when intraday volatility is high and the ob-
served effective spread is low. We use the Best Bid Offer (BBO) dataset from the electronic
trading system CME Globex for corn futures. The data spans 2008-2011 and 2013-2016.8
The TAQ data manipulation protocol follows Section V in Corwin and Schultz (2012) as
close 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 correspondent trade receives

---
8I am indebted to Quanbiao Shang who kindly made the BBO dataset available.
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 novel 
bid-ask couples, 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 - the minimum tick -, and prices are always from the 
nearby contract. The rolling procedure is the first day of the expiration month, and we also 
recompute CSHL estimates for daily data under the same maturity rollover choice.

In the BBO dataset, it is usual to observe multiple quotes for the same time-stamp. 
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 follows: 
\[ e_i \equiv 2 |P_i - M_i|/M_i, \]  
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 
bid-ask spread, BAS, is simply the difference between ask and spread quotes at every second. 
We average BAS over the trading day to obtain daily BAS values. Lastly, intraday volatility 
is measured by the standard deviation of \( M_i \) across all seconds in a given day. Figures (4.5) 
and (4.6) depict the dollar dimension of bid-ask spreads in corn futures.

In Figure (4.4), we subsample the corn futures data so that days corresponding to cases 
2 and 3 are excluded. The idea is to consider days when the CSHL can only be positive, 
by refining the standard negativity adjustment from Corwin and Schultz (2012) which also 
incorporates cases 2 and 3 when spreads turned out positive. Because in practice we are 
selecting even fewer days than the positive-only adjustment, measurement error, given by 
CSHL \( - S^* \), where we assume \( S^* \) to be the effective spread, is consistent and always positive. 
Therefore, even when we exclude cases where negativity can arise, which amounts to 22\% 
of trading days, the CSHL estimator still suffers from measurement error and significantly 
overestimates effective spreads at the daily level.

Table (4.2) presents the main results comparing the microstructure data spreads and 
high-low estimates. We observe an overall pattern of overestimation of the CSHL measure 
when the negative adjustment is included for all years. Averaging over the full sample yields 
slightly better results for years such as 2013, and close estimates in 2014 and 2015. However, 
some averages are negative, which makes it difficult to justify choosing the full sample average 
as the aggregation method. The frequency of negative spreads stays around 40\%, a level
which indicates high measurement error in our experimental setting. As mentioned before, averaging over only positive spreads represents an upper-bound for estimates (with respect to the full sample and the negativity adjustment), hence overestimation becomes even more problematic. In the terminology introduced, positive spreads correspond to Cases 1 and 4, all minor cases, and Cases 2 and 3 that did not result in negative estimates.

The figure displays the difference CSHL−$S^*$, where $S^*$ is the effective spread obtained as explained in the main text. The 350 days selected only include Cases 1, 4, A, B, C, D, and E, where the high-low spread is always positive-definite.

Figure 4.4: CSHL estimator measurement error

As a final comparison measure, Table (4.3) reports annual correlation coefficients for CSHL estimates and effective spreads. The results are clearly unfavorable to support the high-low spread estimator as a good proxy for trading costs in corn futures, specially with the recommended ad hoc adjustment for negative spreads, where displayed negative coefficients align with findings in Marshall et al. (2015). In the previous subsection, we established conditions for which the CSHL estimator is negative. We argued that negative estimates could be related to the measurement error direction, which in turn could be used to assess how much error CSHL estimation makes one incur into. In a market characterized by small spreads, the high-low estimator performs poorly even when it can only be positive. However, Table (4.3) also shows some general improvement in correlation terms between the spread benchmark and CSHL when potential negative spreads (cases 2 and 3) are dropped. For example, in 2016, a correlation coefficient of almost 0.25 represents an improvement with
respect to the correlation of 0.03 when the negativity adjustment is implemented. Notwithstanding, abnormal coefficients in 2011, 2013 and 2015 still make the CSHL unreliable.

Figure 4.5: Daily average bid-ask spreads for corn futures: 2008-2011

Figure 4.6: Daily average bid-ask spreads for corn futures: 2013-2016
Table 4.2: CHSL estimates and actual trading costs in corn futures

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective Spread (%)</td>
<td>0.092</td>
<td>0.095</td>
<td>0.093</td>
<td>0.062</td>
<td>0.061</td>
<td>0.065</td>
<td>0.070</td>
<td>0.070</td>
</tr>
<tr>
<td>BAS</td>
<td>0.300</td>
<td>0.269</td>
<td>0.264</td>
<td>0.281</td>
<td>0.255</td>
<td>0.263</td>
<td>0.260</td>
<td>0.258</td>
</tr>
<tr>
<td>Volatility</td>
<td>4.516</td>
<td>2.814</td>
<td>2.743</td>
<td>4.891</td>
<td>1.520</td>
<td>2.051</td>
<td>1.862</td>
<td>1.259</td>
</tr>
</tbody>
</table>

CSHL (%)

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative adjustment</td>
<td>0.720</td>
<td>0.602</td>
<td>0.519</td>
<td>0.599</td>
<td>0.415</td>
<td>0.529</td>
<td>0.511</td>
<td>0.406</td>
</tr>
<tr>
<td>Full sample</td>
<td>-0.199</td>
<td>-0.086</td>
<td>0.006</td>
<td>-0.047</td>
<td>0.181</td>
<td>0.089</td>
<td>0.091</td>
<td>0.131</td>
</tr>
<tr>
<td>Positive only</td>
<td>1.284</td>
<td>1.059</td>
<td>0.883</td>
<td>1.061</td>
<td>0.671</td>
<td>0.894</td>
<td>0.810</td>
<td>0.590</td>
</tr>
<tr>
<td>Without cases 2 and 3</td>
<td>1.375</td>
<td>1.253</td>
<td>0.902</td>
<td>1.223</td>
<td>0.769</td>
<td>1.095</td>
<td>1.136</td>
<td>0.805</td>
</tr>
<tr>
<td>Frequency negative (%)</td>
<td>43.95</td>
<td>43.15</td>
<td>41.20</td>
<td>43.55</td>
<td>38.10</td>
<td>40.80</td>
<td>36.90</td>
<td>31.08</td>
</tr>
</tbody>
</table>

Number of days

|       | 244  | 252  | 252  | 252  | 21   | 250  | 252  | 74   |

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, Garcia, and Irwin (2014), also in cents/bushel. The high-low spread estimates encompass the same samples used in the high-frequency benchmarks, with rollover on the first-day of the expiration month. The yearly value displayed for all variables represents the simple average over all daily observations within the correspondent year.

Table 4.3: Correlation of CSHL estimates and high-frequency benchmark

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative adjustment</td>
<td>-0.076</td>
<td>-0.038</td>
<td>-0.047</td>
<td>0.002</td>
<td>0.006</td>
<td>0.015</td>
<td>-0.015</td>
<td>0.033</td>
</tr>
<tr>
<td>Full sample</td>
<td>-0.173</td>
<td>-0.126</td>
<td>-0.031</td>
<td>0.011</td>
<td>0.034</td>
<td>0.051</td>
<td>0.024</td>
<td>-0.071</td>
</tr>
<tr>
<td>Without cases 2 and 3</td>
<td>0.126</td>
<td>0.137</td>
<td>0.120</td>
<td>0.034</td>
<td>0.033</td>
<td>0.228</td>
<td>-0.145</td>
<td>0.242</td>
</tr>
</tbody>
</table>

Notes: Yearly correlations of high-low estimates with effective spreads.
The results in this section demonstrate in the context of commodity futures that the CSHL measure inherently suffers from lack of precision, which arises from two sources. First, greater volatility widens estimates so that the absolute value of measurement error increases at daily level. When spreads are small, a large portion of underestimates becomes negative. Negativity is not only easily achieved, but marginal price changes away from the function \( G \) at \( G = 1 \) make negative values even more negative. That explains why averaging over the full sample usually lead to negative estimates, as the values of negative spreads exceeds overstated estimates. On the other hand, no combination resorting to only positive spreads yields good proxies, because the positive measurement error is always large.
Chapter 5

CONCLUSIONS

As new estimators to infer transaction costs in financial markets become available, assessing intrinsic properties, empirical stability, data requirements for implementation and meaningfulness across different market microstructures is fundamental to determine whether a given estimator can safely proxy trading costs when these are not available. Such circumstances include obtaining historical spread estimates and applications for markets which still lack high-frequency data available. Most spread estimators are derived under equity trading, whose microstructure is different from commodity trading. Thus, it is plausible that good performance results in equity venues will not extend to commodity markets. In this work we evaluate the performance of one of such estimators, the high-low spread measure, under a number of conditions, to determine whether the proxy constitute a good alternative for estimating trading costs in commodity markets.

We start by showing historical spread estimates for corn, soybeans, wheat, sugar, live cattle and Brent oil futures contradict economic intuition that trading costs have decreased over time. That motivates a simulation environment where we show the CSHL to yield large measurement error, especially when the volatility-to-spread ratio is not too small. Measurement error amounts to as much as 16-fold when the true spread is below 1%. Since the frequency of negative estimates also increases as volatility gets larger and true spreads smaller, we investigate the determinants of nonpositivity in the CSHL estimator. Negative values may arise with surprising facility, and may be connected to the following dynamics. The CSHL inherently suffers from lack of precision, overestimating and underestimating spreads for all levels of trading costs. As volatility increases, measurement error widens as a consequence of estimates moving even further away from the true spread level. When the spread is sufficiently small, many understated spreads will turn out negative, such that no averaging process delivers good average proxies.

We confirm the findings suggested in the simulation framework with trade-and-quote data from 2008-2016. Comparisons between the effective spread and CSHL estimates indicate
a large positive measurement error of the high-low estimator even when only cases that can only yield positive estimates are considered. Based on two main findings, namely, the poor performance of the CSHL estimator when true spreads are small, as with commodity markets in general, and the necessity of a priori knowledge of the true spread to determine the error size from employing the estimator as a proxy, we do not recommend the use of the high-low spread estimator in commodity markets. Furthermore, as the actual performance of CSHL is conditioned to the true spread size, we suggest the high-low measure not to be a well-behaved estimator in a more general sense. The results of this paper can benefit two different research agendas. First, we confirm previous results for other spread estimators that perform well in equity markets fail to properly represent transaction costs in commodity markets. Unfortunately, an estimator such as CSHL which improved on other measures for equity markets, remains inappropriate to construct historical spreads which allow policy makers, for example, to assess how market efficiency in commodity trading has changed over time. This suggests that developing spread estimators that better represent commodity markets remains a challenge. In addition, this work sheds light on more delicate issues concerning the high-low spread that fundamentally affect its performance in any market with a sufficiently large volatility-to-spread ratio. Whether this performance issue may be solved or at least ameliorated introduces novel directions for further developing the high-low estimator. We hope to deal with the latter in another paper.


APPENDIX A: Simplified form of $\alpha$

Let $\alpha$ be a function of $\beta$ and $\gamma$ without alluding to each parameter’s own arguments. The full CSHL formula for $\alpha$ is given by

$$\alpha(\beta, \gamma) = \sqrt{2\beta - \beta} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}}.$$  \hfill (6.1)

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

$$\alpha = \frac{\sqrt{\varphi}\left[\sqrt{\beta}\left(\sqrt{2} - 1\right) - \sqrt{\gamma}\sqrt{\varphi}\right]}{\varphi\sqrt{\varphi}} = \frac{\sqrt{\beta}\left(\sqrt{2} - 1\right) - \sqrt{\gamma}\left(\sqrt{3 - 2\sqrt{2}}\right)}{3 - 2\sqrt{2}}$$  \hfill (6.2)

which can be further simplified into

$$\alpha = \frac{(\sqrt{2} - 1)\left(\sqrt{\beta} - \sqrt{\gamma}\right)}{3 - 2\sqrt{2}}$$  \hfill (6.3)

since $\sqrt{3 - 2\sqrt{2}} = \sqrt{2} - 1$. By employing a similar procedure, we arrive at the simplified version of the CSHL argument, $\alpha$:

$$\alpha = \frac{(3 + 2\sqrt{2})\left(\sqrt{2} - 1\right)\left(\sqrt{\beta} - \sqrt{\gamma}\right)}{(3 + 2\sqrt{2})(3 - 2\sqrt{2})} = (\sqrt{2} + 1)\left(\sqrt{\beta} - \sqrt{\gamma}\right).$$  \hfill (6.4)
APPENDIX B: Strict positivity of minor cases

Case A: Case A is defined as $H_1 > H_2$ and $L_1 = L_2$. Call the latter $a$, which combined with the former defines the following NC:

$$
\left( \ln \left( \frac{H_1}{a} \right) \right)^2 > \left( \ln \left( \frac{H_1}{a} \right) \right)^2 + \left( \ln \left( \frac{H_2}{a} \right) \right)^2
$$

(7.5)

which cannot hold since $H_2 > a$.

Case B: Here, $H_1 = H_2 = b$ and $L_1 < L_2$. Thus,

$$
\left( \ln \left( \frac{b}{L_1} \right) \right)^2 > \left( \ln \left( \frac{b}{L_1} \right) \right)^2 + \left( \ln \left( \frac{b}{L_2} \right) \right)^2
$$

(7.6)

which is also unfeasible.

Case C: Now, $H_1 = H_2 = b$ and $L_1 = L_2 = a$. It is straightforward to see that $\gamma = 2\beta$, $\beta > 0$. Therefore, the spread is always strictly positive.

Case D: $L_1 = L_2 = a$, but $H_1 < H_2$. This is symmetric to Case A and by the same argument, the negativity condition cannot be sustained.

Case E: Finally, this case is symmetric to Case B and similarly, the spread is always positive.