New Bid-Ask Spread Estimators from Daily High and Low Prices

Zhiyong Li*, Brendan Lambe, Emmanuel Adegbite

Department of Accounting and Finance, De Montfort University, UK

September 2017

Abstract

Estimating trading costs in the absence of recorded data is a problem that continues to puzzle financial market researchers. We address this challenge by introducing two low frequency bid-ask spread estimators using daily high and low transaction prices. The range of mid-prices is an increasing function of the sampling interval, while the bid-ask spread and the relationship between trading direction and the mid-price are not constrained by it and are therefore independent. Monte Carlo simulations and data analysis from the equity and foreign exchange markets demonstrate that these models significantly out-perform the most widely used low-frequency estimators, such as those proposed in Corwin and Schultz (2012) and most recently in Abdi and Ranaldo (2017). We illustrate how our models can be applied to deduce historical market liquidity in US, UK, Hong Kong and the Thai stock markets. Our estimator can also effectively act as a gauge for market volatility and as a measure of liquidity risk in asset pricing.

Keywords: High-low spread estimator; effective spread; transaction cost; market liquidity

JEL Classification: C02, C13, C15

*Corresponding author. Email: Zhiyong.Li@dmu.ac.uk. Address: Hugh Aston 1.79, De Montfort University, The Gateway, Leicester, LE1 9BH

1 Introduction

Estimating trading costs in the absence of recorded data is a problem that continues to puzzle financial market researchers. A bid-ask spread estimator that can be used on a range of market instruments with minimum input data and which is both accurate and efficient has become the *sine qua non* in scholarly research when true trading cost data is unavailable. Much effort is spent on arriving at an estimator that resolves this issue of opacity in historical trading cost data. New models are being introduced and ideas on how to estimate spreads have evolved considerably in the years since Roll (1984) originally formulated the concept. In this paper, we introduce two low frequency bid-ask spread estimators, these can create estimates of costs using the range between daily and two day high and low prices.

We also demonstrate how the models we propose significantly outperform existing versions, namely those introduced by Corwin and Schultz (2012) (the CS estimator hereafter), Abdi and Ranaldo (2017) (the AR estimator hereafter) and the benchmark estimator introduced in Roll (1984). We perform tests using Monte Carlo simulations, and real foreign exchange and U.S. equity market data. Our models are designed along similar principles to the CS estimator in that it assumes that high and low prices are based on buy and sell transactions respectively. We present two models, the first, which we call our basic version uses a transaction range which is determined in part by the mid-price range and by the bid-ask spread. We posit that the mid-price range is a function of the time interval from which it is calculated. Therefore, by comparing the ranges of transaction prices from two different sampling frequencies, we can isolate the impact of the bid-ask spread. Our second model, which we refer to as our sophisticated version, builds on ideas proposed in Bleaney and Li (2016) (the BL estimator hereafter). In the model that we present, the bias that occurs as a result of feedback trading which is evident in the BL model is used to link the one and two day ranges in order to arrive at an estimation of the spread. We note that the bias that results through feedback trading is a function of the time interval. By comparing both the one and two day BL spreads we obtain our estimates of the bid-ask spread. In order to analyse estimator performance we examine the mean and standard deviation of estimated errors alongside the correlation between those and the true spreads. In addition we move beyond simply relying on one single criterion such as correlation to indicate performance. We instead show that to gauge performance on a range of indicators, such as mean, standard deviation and root mean square error (RMSE), is the optimum path to choose for researchers who are keen to attain a more accurate measure of trading costs.

The estimation of accurate bid-ask spreads has for a long time been considered a significantly important part of market microstructure theory. Bid-ask spread estimators allow researchers and practitioners to develop trading strategies that incorporate an idea of the costs attached to each transaction. In turn, this allows for a more accurate determination, of the profitability that follows on from applying . Understanding market liquidity is also important to researchers, so a precise estimation of the bid-ask spread offers a clearer picture of this market characteristic (Mancini et al. 2013; Banti et al. 2012). Another possible use for these models arises from the fact that bid-ask spreads can influence measures associated with price volatility, so scholars analysing this metric can use estimator models to arrive at an accurate measure of this (e.g. Bandi and Russell 2006).

As excessive costs are attached to accessing bid-ask spread data this has meant that researchers increasingly rely on such estimators to aid their analysis of market activity. Much research supports this approach indicating that the cheaper daily closing quoted bid-ask spread can be good proxy for the intraday spread (Holden and Jacobsen 2014, Chung and Zhang 2014 and Fong et al. 2017). Inavailability of spread data is not simply a consequence of poor research budgets, historical information on both quoted and true spreads is not always available, a strong performing estimator model is useful even to well-resourced researchers.

Bid-ask spread estimation models need to satisfy certain requirements before they become useful to researchers. Models must be accurate, efficient in terms of having a low standard deviation of estimates and it is preferable that they have low requirements on the type of data needed for computation. In order to improve on accuracy and efficiency, the signal to noise ratio becomes an important consideration; this is because the spread (signal) is more difficult to estimate when it is considerably smaller than the mid-price volatility levels (noise). Assets with higher levels of liquidity demonstrate typically smaller spreads; however with more infrequently traded instruments, the bid-ask spreads can be quite large. Longer sampling intervals have a tendency to display higher mid-price volatility levels, therefore, the signal to noise ratio is smaller in more infrequent samples. This leads to poorer performance in the accuracy and efficiency of estimators that rely on low sampling frequencies, this is pointed out by Bleaney and Li (2015). In addition to the need for models to be both accurate and efficient, other barriers to inquiry may inhibit a model's usefulness. Constraints on accessing data imposed by availability or cost mean that models with more modest data requirements are of greater use to researchers. For instance, Roll (1984) requires just the transaction price of assets in order to apply the estimator, whereas Huang and Stoll (1997) require both the transaction price and the trade direction. Corwin and Schultz (2012) require the high low price range. Abdi and Ranaldo (2017) require the closing price and the high-low price range.

In this paper, we analyse the performance of the estimators through conducting a series of tests using both randomly generated and real data, the latter is taken from both the foreign exchange and equity markets. In most cases, both of our estimators outperform all others tested. In comparison, the CS estimator exhibits instability as it only works well for equities. The AR estimator produces estimates which are highly correlated with the true spread, while displaying a tendency to remain lower than those we estimate, it also performs poorly in terms of the root mean square error (RMSE). Simulation experiments produce a signal to noise ratio over 125000 months of generated data ranging from 0.005 to 0.387. This covers most cases which have occurred in actuality. Both of our proposed models outperform the others we test in both efficiency and accuracy. We also move beyond time series testing to investigate the cross sectional performance on a generated sample of 75000 data months, again at this level we find that our models outperform the others tested.

In the existing literature, bid-ask spread estimators are tested using the price data taken from the equities markets. An additional benefit offered by our models is that these are suitable for use both in the equities and foreign exchange markets because they are independent of the market structure. In this paper we run tests using data taken from both types of financial markets. In both markets the empirical tests are conducted using spreads calculated from tick by tick data as a benchmark, we find that our estimators again perform better than the other models currently available to researchers.

In presenting the significant contributions of our work, the rest of our paper is organised as follows. Section 2 discusses existing bid-ask spread estimators, while section 3 introduces our new models. In section 4 the performances of these are reported against that of the Roll, CS and AR estimators. Section 5 provides an illustrate of some applications of our estimator using equity markets while section 6 concludes.

2 Relevant bid-ask spread estimators

Spread estimator models are generally classified into one of four categories, the Roll, the LOT, the Effective Tick and the more recent High-low estimator. Each approach provides alternative methods which are based on the return autocovariance, the interval fractions in trade prices, the frequency of zero returns and the specific interval determined price range.

Roll (1984) was the first to propose a bid-ask spread estimator. This model was popularly received generating considerable interest at the time, giving rise to attempts to refine it further later by other scholars. Roll's premise was to use return autocovariance to estimate the spread. Underpinning this approach was the assumption that prices followed a random walk. It was also assumed that the closing stock price equalled its true value plus or minus half of the effective spread. The estimated spread could then be calculated as twice the square root of minus one multiplied by the autocovariance of the sample of daily returns. Some problems with this approach have been noted, for instance, the estimator produces results which can often underestimate the spread (Harris 1990). To deal with this autocorrelated mid-price return bias[a], George et al. (1991) suggest modifying the original Roll estimator. Similarly, Choi et al. (1988) introduce adjustments to the model in an attempt to deal with the problem of auto correlated trade directions. Stoll (1989) tackles the problem by taking the impact of inventory control and asymmetric information costs into account. To reach a general solution to the problem, Huang and Stoll (1997) incorporate each of the estimators above in one general model. However, gathering the data required to run this is a difficult process as trade direction data is also required. Hasbrouck (2004, 2009) suggests that a more accurate spread can be achieved through employing Gibbs estimation. Unlike

[a]The bias arises as the assumption that returns are random is not satisfied.

Roll's approach the computational requirements to employ Hasbrouck's estimators are considerably more intensive. The problem of normality in Hasbrouck's is addressed in Chen et al. (2016) who propose a non-parametric method to estimate the spread based on the Roll model. A further development is found in Abdi and Ranaldo (2017) who incorporate the CS model into the Roll to derive a new estimator. The correlation between its estimates and the true spread is higher than the CS estimator, but the RMSE is not significantly better than the CS.

Another estimator used to deal with the problem of the spread opacity due to the inavailability of data is proposed by Lesmond et al. (1999). Otherwise known as the LOT model, an effective spread is calculated by considering the fraction of returns which are different from zero. This model has not quite reached the popularity levels of the Roll model as comparatively it tends not to perform as well in empirical testing (Corwin and Schultz 2012). Holden (2009) and Goyenko et al. (2009) put forward a more sophisticated approach which estimates spreads using effective tick measures based on the phenomenon of price clustering, a term describing the tendency for trade prices to occur most frequently on rounder price increments. However, results produced following testing on an extensive FX market data sample by Karnaukh et al. (2015) show that the LOT and effective tick estimators display only a weak relationship with true spreads.

The high-low spread estimator introduced by Corwin and Schultz (2012) adds new power to the toolkit of estimators. Despite being relatively new, it is used extensively in recent literature as testing shows that it satisfies the estimator requirements to a greater extent than previous innovations (Corwin and Schultz 2012, Holden and Jacobsen 2014 and Karnaukh et al. 2015). The model is derived using the high and low prices of an asset over two day and daily horizons. The relative ease with which input data can be accessed further enhances the appeal of the estimator. The estimator we introduce here operates in the spirit of the CS model by adopting similar assumptions that the high (low) prices are most likely buy (sell) orders.

3 The High-low estimators

The general structure of both of our estimators can be expressed in the following equation:

$$\frac{E\left(\sqrt{2} \cdot X_{daily} - X_{twoday}\right)}{\sqrt{2} - 1} \tag{1}$$

Where X is the price range or the estimated spread by the Bleaney and Li (2016) estimator. The innovations we introduce for both the basic and sophisticated models are discussed in the following subsections.

3.1 The basic high-low estimator (the BHL model)

The basic high-low estimator is similar to the CS model, but unlike its quadratic structure, the BHL uses a linear structure which assures the unbiasedness and consistency of its estimates. Furthermore, the combination of BHL and the sophisticated high-low estimator introduced in the next section sometimes works better than a single estimator, because the estimated errors from different estimators offsets each other to some extent.

When estimating a model using the high and low transaction prices, the first characteristic that we can note is that the range increases as the time interval widens and as the mid-price volatility grows in proportion. The other factor contributing to the range is the bid-ask spread, however this is independent of the time interval. Therefore, it is possible to extract the bid-ask spread by calculating the difference between the high and low transaction prices, whilst considering inconsistencies that may arise as a result of volatility.

In order to accomplish this, we assume that the mid-price, denoted as M_t, follows a one-dimensional Wiener process. The link between the unobserved mid-price and the observed transaction price (s_t) is given through the following equation.

$$s_t = M_t + \frac{SP}{2} \cdot BS_t \tag{2}$$

Where BS_t is the trade indicator showing 1 (-1) for a buyer (seller) initiated trade. The relationship between the daily high mid-price (H_t^M) and the daily high transaction price (H_t^T) as well as the link between the daily low mid-price (L_t^M) and the daily low transaction price (L_t^T) are demonstrated in the following set of equations:

$$\begin{array}{ll} H_t^T = H_t^M + \frac{SP}{2} \cdot BS_t & L_t^T = L_t^M + \frac{SP}{2} \cdot BS_t \\ TH_t^T = TH_t^M + \frac{SP}{2} \cdot BS_t & TL_t^T = TL_t^M + \frac{SP}{2} \cdot BS_t \end{array} \tag{3}$$

Where T and M represent the transaction and mid-price respectively. TH and TL denote the high and low prices over a two day window.

We can eliminate the need to establish trade direction by assuming that the highest (lowest) prices are buy (sell) orders. Formally, it can be represented as:

$$\mathrm{BS_t} = \begin{cases} 1 \quad if \;\; \mathrm{s_t} = H_t^T \\ -1 \;\; if \;\; \mathrm{s_t} = L_t^T \end{cases} \tag{4}$$

The daily and two-day ranges of transaction prices represent the difference between the highest and lowest prices. Formally, taking equations (3) and (4) into account, these ranges are given as:

$$\begin{aligned} Range_{t,daily}^T &= H_t^T - L_t^T \\ &= \left(H_t^M + \tfrac{SP}{2} \cdot BS_t\right) - \left(L_t^M + \tfrac{SP}{2} \cdot BS_t\right) \\ &= \left(H_t^M + \tfrac{SP}{2}\right) - \left(L_t^M - \tfrac{SP}{2}\right) \\ &= (H_t^M - L_t^M) + SP \\ &= Range_{t,daily}^M + SP \end{aligned} \tag{5}$$

$$\begin{aligned} Range_{t,twoday}^T &= TH_t^T - TL_t^T \\ &= \left(TH_t^M + \tfrac{SP}{2} \cdot BS_t\right) - \left(TL_t^M + \tfrac{SP}{2} \cdot BS_t\right) \\ &= \left(TH_t^M + \tfrac{SP}{2}\right) - \left(TL_t^M - \tfrac{SP}{2}\right) \\ &= \left(TH_t^M - TL_t^M\right) + SP \\ &= Range_{t,twoday}^M + SP \end{aligned} \tag{6}$$

Where $Range_{t,daily}^T$ and $Range_{t,twoday}^T$ are daily and two-day ranges respectively. The equations above demonstrate our earlier suggestion that the range of transaction prices is influenced by volatility in both the mid-price and the bid-ask spread. Taking expectations of both sides, the equations become:

$$E\left(Range_{daily}^T\right) = E\left(Range_{daily}^M\right) + SP \tag{7}$$

$$E\left(Range_{twoday}^T\right) = E\left(Range_{twoday}^M\right) + SP \tag{8}$$

The left hand sides of Equations (7) and (8) can be calculated from observed transaction prices. With the unobserved terms, the expected ranges of daily and two-day mid-prices can be eliminated, allowing us to extract the bid-ask spread. Parkinson (1980) shows that if the mid-price follows a one-dimensional Wiener process, its expected range is an increasing function of the sampling time interval and its diffusion. A long sampling time interval or large diffusion will lead to a wider range. Formally, the expectation of the range of mid-prices can be calculated through the following equation:

$$E\left(Range^{M}\right) = \sqrt{\frac{8D \cdot ti}{\pi}} \tag{9}$$

Where D is the diffusion of mid-prices in a unit time interval (ti). If this period is one day, the expectations for daily and two-day ranges are given through the following equations:

$$E\left(Range^{M}_{daily}\right) = \sqrt{\frac{8D}{\pi}} \tag{10}$$

$$E\left(Range^{M}_{twoday}\right) = \sqrt{\frac{8D}{\pi}} \cdot \sqrt{2} \tag{11}$$

Therefore, the expectation is that the two-day range is $\sqrt{2}$ times that of the daily range. Formally, the relationship is expressed through the following equation:

$$E\left(Range^{M}_{twoday}\right) = \sqrt{2} \cdot E\left(Range^{M}_{daily}\right) \tag{12}$$

From Equations (7), (8) and (12), we can solve for the bid-ask spread (SP), because we have three equations and three unknown variables. We solve Equation (8) through deducting $\sqrt{2}$ times each side of Equation (7):

$$\begin{aligned} &E\left(Range^{T}_{twoday}\right) - \sqrt{2} \cdot E\left(Range^{T}_{daily}\right) \\ &= E\left(Range^{M}_{twoday}\right) + SP - \sqrt{2} \cdot \left[E\left(Range^{M}_{daily}\right) + SP\right] \end{aligned} \tag{13}$$

When we substitute Equation (12) into (13), and rearrange the yields, the estimate of the bid-ask spread becomes:

$$SP = \frac{E\left[\sqrt{2} \cdot \left(Range^{T}_{daily}\right) - \left(Range^{T}_{twoday}\right)\right]}{\left(\sqrt{2} - 1\right)} \tag{14}$$

Equation (14) represents the basic estimator which we propose in this paper (BHL hereafter); this is an expectation of the linear function of the daily and two-day high and low transaction prices. One of its key features is that it is unbiased and easy to compute. It outperforms the CS estimator because it produces an unbiased result while remaining linear. Using BHL, it is possible to increase the number of observations in order to obtain a more accurate and efficient estimate of the spread. The reason is that statistical errors and noise can be eliminated from large sample sizes; this is not the case for non-linear estimators (Bleaney and Li 2015). Furthermore, the estimates remain consistent across a variety of sampling periods. This suggests that when higher sampling frequency data becomes available we can use it to obtain a more accurate and efficient estimate because, as Bleaney and Li (2015) suggest, the noise (the price volatility) is relatively low in comparison with the bid-ask spread.

Similar to the CS estimator, we can also estimate the daily diffusion, which is expressed as D, using the same process. This is represented in the following equation:

$$\begin{aligned} E\left(Range^{M}_{daily}\right) &= \frac{E\left(Range^{T}_{twoday}\right)-E\left(Range^{T}_{daily}\right)}{\left(\sqrt{2}-1\right)} = \sqrt{\frac{8D}{\pi}} \\ D &= \frac{\pi}{8}\left[\frac{E\left(Range^{T}_{twoday}\right)-E\left(Range^{T}_{daily}\right)}{\left(\sqrt{2}-1\right)}\right]^{2} \end{aligned} \tag{15}$$

3.2 The sophisticated high-low estimator

Our sophisticated estimator (the SHL model hereafter) introduces innovations to the design proposed by Bleaney and Li (2016). The BL estimator is distinctive in that it outperforms Roll (1984), Huang and Stoll (1997), Corwin and Schultz (2012) and Hasbrouck (2009) estimators, following extensive testing. The disadvantage with the BL model is that its computation requires both the transaction price and information on the direction of trade and often researchers don't have access to this information.

SHL introduces the assumption that the highest prices recorded daily are ask-prices and the lowest are bid-prices. Through this assumption we can lower the data requirements for the model and allow the estimator to operate using only the highest and lowest transaction prices in the estimation window.

In a similar manner to Bleaney and Li (2016), we assume that we have random conjectures of the true bid-ask spread. We let set A be a set of all conjectures where the symbol $\sim$ represents conjectural values.

$$A = \left\{\widetilde{SP}_1, \widetilde{SP}_2, \cdots, \widetilde{SP}_n\right\} \tag{16}$$

At this stage, we do not know which element in set A is the true spread. Through taking the following steps, we would be able to find it. First, we would calculate a series of conjectural mid-price returns according to each element (a conjectural spread) in set A using equation (2). Formally, the conjectural mid-price return is given as follows,

$$\widetilde{M}_t = s_t - \frac{\widetilde{SP}_t}{2} \cdot BS_t \tag{17}$$

Second, we calculate the variance of conjectural mid-price returns for each conjectural series.

B denotes a set of variances of conjectural mid-price returns, the conjecture being that the true spread is taken to be:

$$B = \{Var_1, Var_2, \cdots, Var_n\} \tag{18}$$

Where

$$Var_i = Var\left[\Delta\widetilde{M}(\widetilde{SP}_i)_t\right] \tag{19}$$

Third, based on these settings we can find the true spread among the conjectures by find the biggest relevant variance. Formally, we propose the following:

Proposition 3.1 *When the components of the spread do not include feedback trading, inventory control or asymmetric information, we can consider that the spread and its estimates, and thus the estimated errors, are either serially independent or fixed. If an estimate of the spread* $\widetilde{SP}_i \in A$ *corresponds to* $Var_i = max(B)$*, it equals the true spread which is then denoted as:* $\widetilde{SP}_i = SP$*.*

Proof The full proof is given in the appendix. The variance of two adjacent conjectures of mid-price returns is:

$$\begin{aligned} Var_i &= Var\left[\Delta\widetilde{M}_t\right] \\ &= E\left\{\left[\Delta\widetilde{M}_t - E\left(\Delta\widetilde{M}_t\right)\right]^2\right\} \end{aligned} \tag{20}$$

We will assume that the expectation of the value of the conjectural mid-price is zero. Thus, the equation above can be rewritten as:

$$\begin{aligned} Var_i &= Var\left(\Delta\widetilde{M}_t\right) \\ &= E\left(\Delta\widetilde{M}_t^{\,2}\right) \\ &= E\left[\left(\Delta\, M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)^2\right] \end{aligned} \tag{21}$$

Where Ω denotes the conjectural error which represents the difference between the conjectural mid-price and the true mid-price, alternately expressed as the difference between the conjectural spread and its true value. Formally, Ω is given as:

$$\Omega = \Delta\widetilde{M}_t - \Delta M_t = \widetilde{SP}_i - SP \tag{22}$$

The assumptions of this proposition imply that BS is independent of $\Delta\, M$ at all observation points, therefore many of the terms in (21) such as $E(\Delta\, M_t\, BS_t)$ and $E(\Delta\, M_{t-1}\, BS_t)$ equate to zero. The variable BS is a binary variable (1 or -1), thus $E\left(BS_{t-1}{}^2\right) = 1$. Finally we obtain:

$$\begin{aligned}
&Var_i = Var\left(\Delta\widetilde{M}_t\right) \\
&= E\left(\Delta\widetilde{M}_t{}^2\right) \\
&= E\left[\left(\Delta\, M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)\left(\Delta\, M_t + \tfrac{1}{2}\Omega\, BS_t - \tfrac{1}{2}\Omega\, BS_{t-1}\right)\right] \\
&= E\left(\Delta M_t{}^2 + \tfrac{1}{2}\,\Omega^2\right)
\end{aligned} \tag{23}$$

The final step of Equation (23) given above is the quadratic polynomial of the expectation of the error of the conjecture. For a given series, the first term $E(\Delta M_t^2)$ is a constant. We can surmise directly from this that when the error is zero (i.e. $\Omega = 0$), the second term $\frac{1}{2}\Omega^2$ is zero. Furthermore, when $\Omega = 0$, there is a global extreme for the right hand side polynomial in the final step, and symmetrically, the left hand side of the equation $Var_i = Var\left(\Delta\widetilde{M}_t\right)$ is also at the extreme value. Formally this can be expressed as:

$$\underset{\Omega}{arg\ max}\ Var\left(\Delta\widetilde{M}_t\right) = 0 \tag{24}$$

When the conjectural error is zero, the conjectural spread becomes the true spread:

$$\widetilde{SP}_i = SP + \Omega = SP \tag{25}$$

Therefore the conjectural spread which maximises the covariance equals the true spread.

$$\underset{\widetilde{SP}_i \in A}{arg\ max}\ Var\left(\Delta\widetilde{M}_t\right) = SP \tag{26}$$

Q.E.D.

Figure 1 outlines the reasoning underpinning this proposition where for the purposes of economy we hold that the mid-price is fixed, , while mid-prices following a random walk will not affect the derivation of the model. The conjectural spread ($\widetilde{SP}_i$) is less than the true spread. This allows us to estimate the conjectural mid-price $\widetilde{M}$; this is represented by the dotted line in Figure 1, and the true mid and transaction prices are both represented by unbroken lines. Also in Figure 1, A and B denote observed ask and bid prices, whereas M is the unobserved true mid-price.

At any one point we can only observe one price, which is either the bid or ask. In Figure 1, three periods are displayed. In the period labelled $t-2$, the bid price is recorded and in period labelled $t-1$, the ask price is observed. In period $t-2$, the conjectural spread is lower than the true spread and the conjectural mid-price error is -0.5Ω , which is less than the true value. In period $t-1$, the conjectural mid-price error is 0.5Ω, therefore this is greater than the true one. In the intervening period between $t-2$ and $t-1$, the direction of the trade shifts from sell to buy, and because of the conjectural error, we overestimate the mid-price return, formally we express this as:

$$\Delta\widetilde{M}_{t-1} = \Delta M_{t-1} + \Omega = \Omega \tag{27}$$

In Figure 1, the hypothetical example shows that the variance of mid-price returns equates to zero because returns remain fixed. However the variance of conjectured mid-price returns is greater than zero. The reason for this is that in the case where the spread is underestimated, the conjectured mid-price fluctuates more than its true counterparts.

[Insert Figure 1 here]

According to the abovementioned proposition, we find that the true spread maximises the variance of conjectural mid-price returns and can be expressed as follows:

$$\begin{aligned}
&Var\left(\Delta\widetilde{M}_t\right) = E\left(\Delta\widetilde{M}_t^{\,2}\right)\\
&= E\left[\left(\Delta\, s_t - \tfrac{\widetilde{SP}}{2}\Delta\, BS_t\right)^2\right]\\
&= E\left(\Delta\, s_t^{\,2}\right) - \widetilde{SP}\cdot E\left(\Delta\, s_t\,\Delta\,\mathrm{BS}_t\right) + \tfrac{\widetilde{SP}^2}{4}\cdot E\left(\Delta\, BS_t^{\,2}\right)
\end{aligned} \tag{28}$$

Using first order conditioning, we find that the estimated spread satisfies the following equation:

$$-\,E\left(\Delta\, s_t\,\Delta\, BS_t\right) + \frac{1}{2}\widetilde{SP}\cdot E\left(\Delta\, BS_t^{\,2}\right) = 0 \tag{29}$$

$$SP = \widetilde{SP} = \frac{2E\left(\Delta s_t \Delta\, BS_t\right)}{E\left(\Delta\, BS_t{}^2\right)} \tag{30}$$

Equation (30) is now the variance version of the BL estimator, thereby reflecting one of the suggested innovations that we propose in this paper.

In order to allow Equation (30) to become operational, we must introduce the following processes. On each day, we pick either the high or low price randomly to create a trail series of prices (s_t) and use Equation (4) to determine the trade direction: a buy order when s_t is the high price and a sell order when s_t is the low price. We can then calculate the estimated spread using Equation (30). In the same manner as Corwin and Schultz (2012), we calculate an estimate of spread using the two-day high and low prices SP_{twoday}.

However, Equation (4) creates the link between order flow and price when only high and low prices are used. When the covariance between order directions and mid-price returns is non-zero, the BL estimator is biased and the error is expressed as $E\left(BS_t \cdot \Delta\, M_t\right)$ [b]. Therefore, when high and low prices and relevant trade directions are used, the BL estimator significantly overestimates the spread. It is invariably the case that the estimated spread will contain errors, these nevertheless can be offset if we compare the estimates using daily and two-day data ($E(\mathrm{SP}_{daily})$ and $E(\mathrm{SP}_{twoday})$). This is because it is possible to predict the relationship between errors from daily and two-day estimates.

The true spread is then taken to be the estimated spread minus the error. Formally, the true spread is given as follows:

$$\begin{aligned} SP &= SP_{daily} - E\left(BS_{daily} \cdot \Delta\, M_{daily}\right) \\ SP &= SP_{twoday} - E\left(BS_{twoday} \cdot \Delta\, M_{twoday}\right) \end{aligned} \tag{31}$$

Where the subscripts "daily" and "twoday" represent the sampling frequencies. SP_{daily} and SP_{twoday} are the BL estimates using daily and two-day high low data respectively. Following the discussion in the previous section, the relationship between daily and two-day ranges can be used to eliminate the estimated error above. The errors are in fact half of expected ranges of daily and two-day ranges and are expressed as follows

[b] The feedback trading bias is discussed in Bleaney and Li (2016). Proofs are given in the appendix of this paper.

(proof can be found in the appendix):

$$\begin{aligned} E\left(BS_{daily} \cdot \Delta\, \mathrm{M}_{daily}\right) &= \tfrac{1}{2} E\left(\mathrm{H}_t - \mathrm{L}_t\right) = \tfrac{1}{2} E\left(Range^{M}_{daily}\right) \\ E\left(BS_{twoday} \cdot \Delta\, M_{twoday}\right) &= \tfrac{1}{2} E\left(TH_t - TL_t\right) = \tfrac{1}{2} E\left(Range^{M}_{twoday}\right) \end{aligned} \tag{32}$$

Following steps similar to the process outlined in section 3.1, we substitute Equations (32) and (12) into Equation (31), the rearrangement yields the spread for one trial series. This trial estimate is given as:

$$SP_{onetrial} = \frac{\sqrt{2}\, SP_{daily} - SP_{twoday}}{\sqrt{2} - 1} \tag{33}$$

We repeat the trail series creation and estimation process over 1000 times, the mean of these estimates becomes our estimation of the spread. Although this process is computationally intensive, this makes little practical difference given the power of the current stock of computers available to researchers.

$$SP = E\left(\frac{\sqrt{2}\, SP_{daily} - SP_{twoday}}{\sqrt{2} - 1}\right) \tag{34}$$

Equation (34) is the sophisticated high low estimator (SHL). Theoretically, SHL should produce more accurate results than its BL counterpart. Unlike the BL model, the SHL estimator will not be influenced by feedback trading and the estimates produced will be unbiased.

When the ratio of the spread to the standard deviation of mid-prices is small, some trail estimates in Equation (33) could be negative. In order to avoid a negative result in Equation (34) we let all negative trial estimates equal zero. However, when we do this, the estimates produced by SHL might overestimate the spread. The simulation experiments we conduct in the next section show that it will not be an issue when the ratio becomes larger.

We can also estimate the daily diffusion, which is expressed as D, from Equations (31) and (32) and using the same process. This is represented in the following equations:

$$\begin{aligned} & SP_{twoday} - SP_{daily} = E\left(BS_{twoday} \cdot \Delta\, M_{twoday}\right) - E\left(BS_{daily} \cdot \Delta\, M_{daily}\right) \\ & = \tfrac{1}{2} E\left(Range^{M}_{twoday}\right) - \tfrac{1}{2} E\left(Range^{M}_{daily}\right) \\ & = \tfrac{(\sqrt{2}-1)}{2} \sqrt{\tfrac{8D}{\pi}} \end{aligned} \tag{35}$$

The rearrangement of the equation above yields an expression of daily diffusion as follows:

$$D = \frac{\pi}{2} \cdot \left[E \left(\frac{SP_{twoday} - SP_{daily}}{\sqrt{2} - 1} \right) \right]^2 \tag{36}$$

4 Comparison of the estimators

In this section, we examine the performance of a range of estimators. Using empirical tests, we gauge how BHL, SHL, Roll, CS and AR perform in addition to a number of equally weighted combinative models. Currently, the range of estimators available to researchers is wide, but we focus on these models for several reasons. The first is that we wish to contrast the performance of our proposed models (BHL and SHL) with that of the best performing estimator available, the CS model. Corwin and Schultz (2012) demonstrate that the CS model outperforms all other low frequency estimators in terms of accuracy and efficiency. Holden and Jacobsen (2014) and Karnaukh et al. (2015) also show similar results to the model originators. We also choose the Roll model as this has traditionally been the benchmark for estimator performance. Researchers less familiar with the relatively recent CS model can understand how our models perform in comparison. Finally, we select the model proposed in Abdi and Ranaldo (2017) because it is the latest development of the spread estimator and is directly related to both the Roll and CS models. Our motivation behind including the combinatory models relates to the tendency for some estimators to over(under)estimate the spread. Combinations of estimators have been shown in (Holden 2009)[c] to perform better in terms of accuracy. The data we use to test each of the models is tick by tick equity prices and foreign exchange rates; these are sourced from TAQ and Hotspot and DataStream respectively. In addition to using real world data testing, simulation experiments were also carried out. Our findings show that in general our BHL and SHL estimators outperform all other estimators included in the study.

[c]Combination models tested in Holden (2009) and used here for testing are explained in Table 1.

4.1 Comparison strategy

Following testing on each of the estimators, the results are reported using average relative estimated errors together with their root mean square and standard deviation values. Formally, the relative error is defined as follows:

$$Rel - Err = \frac{Estimates - Spread}{Spread} \tag{37}$$

The average relative error (Rel-Error-Mean) reports the mean difference between the estimated and true spread, indicating where possible bias may exist in the estimators. When Rel-Error-Mean is positive (negative) it suggests that the models over (under) estimate the spread. Good estimates are those with 'close-to-zero' relative error averages. Formally, this is presented as:

$$Rel - Err - Mean = E\left(Rel - Err\right) \tag{38}$$

The standard deviation of the relative estimated errors (Rel-Err-Std) is also reported and provides a measure for the efficiency of the estimates. Good estimates have low Rel-Err-Stds. Formally, this is expressed as:

$$Rel - Err - Std = Std.Dev\left(Rel - Err\right) \tag{39}$$

Finally, the RMSE is the most widely used criteria by which to judge the performance of the estimators. Therefore we follow this trend in analysing how our models perform. Formally, RMSE is given as:

$$RMSE = \sqrt{E\left[\left(Rel - Err\right)^2\right]} \tag{40}$$

4.2 Simulation experiments

In this section we report the results of a number of simulations designed to test the relative strength of each measure. We find that the estimators proposed in this paper outperform the other models in terms of accuracy and efficiency. Simulation experiments are widely used in literature to examine and to compare various estimators (e.g. Corwin and Schultz 2012, Bleaney and Li 2015, 2016,Karnaukh et al. 2015,Abdi and Ranaldo 2017). Compared to the real data, the statistical properties of estimators can be extracted using a large number simulation experiments. One also could identify the factors that influence the performance of estimators, which help researchers to choose from various estimators according to their purpose.

4.2.1 Estimation under various 'signal to noise' ratios

It is difficult for an estimator to isolate the bid-ask spread from transaction prices when the volatility of mid-prices is relatively large. We test the performance of the estimators under various 'signal to noise' values which are the ratios of the spread to the standard deviation of the mid-price. The 'signal to noise' ratio is low for heavily traded equities and major currency pairs because the liquidity levels are consistently high, and the assumption is that mid-prices and order directions are random. We allow the standard deviation of one-minute mid-price returns to be 0.005 (about 0.19 daily). We consider six bid-ask spreads ranging from 0.001 to 0.3. The 'signal to noise' ratio extends from 0.00527 to 1.58 on a daily basis. In comparison, Corwin and Schultz (2012) test their model using the ratios which begin at 0.167 and end at 3.33; therefore the performance hurdles we employ to evaluate our estimators are more difficult to overcome.

Our simulation experiments are therefore more challenging and mirror real market conditions. For example, assuming that there are 20 trading days in a month, we compare the estimates of 25000 months. Formally, the data generation system is given as:

$$
\begin{aligned}
&s_t = M_t + \tfrac{SP}{2} \cdot BS_t \\
&BS_t \sim B(1, 0.5) \\
&\Delta M_t \sim N(0, 0.05)\ (one - minute) \\
&SP = \begin{cases} 0.001 \\ 0.002\ report\ online\ for\ brevity \\ 0.006\ report\ online\ for\ brevity \\ 0.010\ report\ online\ for\ brevity \\ 0.030 \end{cases}
\end{aligned}
$$

Tables 1 to 5 report, the testing using various versions of our BHL, SHL and CS models[d]. In general, those estimates are more accurate and efficient from the top left to the bottom right, as the ratio (*True spread/Midstd*) increases (from 0.00527 to 0.387) the

[d]See the caption of Table 1 for full details of the versions of the estimators used.

number of observations increases[e]; this is consistent with findings of Bleaney and Li (2015, 2016). By setting negative trials and results to be zero, the BH3, SHL2, CS3, AR and Roll estimators demonstrate significant bias. For example, the relative error of SHL2 is 74.97% and those of BH3, CS3 and AR are 330%, 234% and 155% respectively when the ratio is 0.158 (The left panel of Table 5). If the ratio is 0.387 (The right panel of Table 5), the relative error of SHL2 is −0.593% which is close to zero, and therefore demonstrates the power of the model. For the other estimators, the relative errors of SHL1, BHL1 and BHL2 are less than 5% when the ratio is greater than 0.0258. According to the second column of Table 3, the average ranking of all simulation experiments in section 4.2.1 suggests that the combination of SHL2 and BHL1 is the best performing estimator. In terms of the performance of single rather than combined estimators, SHL2 offers the best results and is the second best performer from the entire array of models.

[Insert Tables 1 to 5 here]

4.2.2 Cross-sectional properties of the estimators

In this section, the cross-sectional properties of the estimators are examined. In contrast to the previous section, the bid-ask spreads are assumed to vary each month and are evenly distributed from 0.002 to 0.0177. We also break the full sample into five groups according to the mean of the bid-ask spread. Thus, we can examine the cross-sectional performances of the estimators across the five ranges of spread. The other parameters in the data generation process are the same as in the previous section.

[e]Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation. We also test the cases of full sample and the case where the trimming is at the 0.05% and 2% level, the results produced are similar.

Formally, the data generation system is given as:

$$
\begin{aligned}
& s_t = M_t + \tfrac{SP}{2} \cdot BS_t \\
& BS_t \sim B(1, 0.5) \\
& \Delta M_t \sim N(0, 0.05)\ (one - minute) \\
& SP = \begin{cases} from\ 0.001\ to\ 0.00513 \\ from\ 0.00513\ to\ 0.00829\ report\ online\ for\ brevity \\ from\ 0.00829\ to\ 0.0114\ report\ online\ for\ brevity \\ from\ 0.0114\ to\ 0.0146\ report\ online\ for\ brevity \\ from\ 0.0146\ to\ 0.0177 \end{cases}
\end{aligned}
\tag{41}
$$

The results of the simulation experiments are reported in Tables 6 to 11. The correlations between the true and estimated spreads are also reported. Table 6 reports the pooled results while the other panels are represented in the equation above according to each grouping of spreads. In the case of spreads where the range is between 0.01 and 0.03, the correlations reported are quite weak.

In the pooled case, although CS3 has a slightly stronger correlation than SHL2 with values of 0.136 and 0.127 respectively, it reports a much higher value for RMSE at 12.32 that the SHL2 value which is 5.51. In terms of correlation, the best performers are CS3, SHL2 and BHL3. From the third column of Table 12, it is evident from the average ranking of all simulation experiments that the combination of SHL2 and BHL1 and that of SHL2 and CS2 are the best performing estimators. For single models, SHL2 shows the best performance and is placed third in rank overall.

Table 12 shows a summarised average ranking for all simulations. According to the first column, the average ranking of all cases of simulation experiments suggests that the combination of SHL2 and BHL1 is the best performing estimator. SHL2 is the best performing single estimator and takes second place overall. The other combinations outperform the other single estimators. The remaining alternate versions of our new models (BHL1, BHL2, SHL1) perform better than all the versions of the CS and Roll estimators.

[Insert Tables 6 to 12 here]

4.3 Comparisons in foreign exchange markets

In this section, we use our chosen estimators to gauge historical spreads for the foreign exchange markets. To do this, we test the estimators using closing prices and quoted spread data of 22 currency pairs in a sample beginning in January 1990 and finishing in December 2016, using data extracted from DataStream. We also use the prices and effective spreads of 12 Currency pairs in a sample dating from December 2015 to August 2016, this data is taken from Hotspot. In testing on both currency samples our sophisticated high-low estimator outperforms all others employed in the test.

4.4 DataStream 22 currency pairs 1990-2016

In this subsection, we evaluate the performance of the estimators using daily high and low prices and closing bid-ask spread data of 22 currency pairs taken from DataStream. For brevity, we report only the pooled results and not the result of the individual tests on each currency pair. The times series properties of the estimators are reported in Table 14 through the average performance ranking of the models for each pair examined; of these, SHL2 and the combination of SHL2 and BHL1 are the best performers.

According to Table 13, the combination of SHL2 and BHL1 outperforms the other estimators in terms of RMSE. Although CS2 has the lowest average estimated error (17.05%), its standard deviation of 2.315 is large relative to the others tested. SHL2 has a lower standard deviation than the aforementioned combination but a greater error (90.49%) and thus it takes second place in terms of performance. CS3 and BHL3 report high correlations between true and estimated spreads with the values of 0.876 and 0.869 respectively. However, the errors and the standard deviations of CS3 and BHL3 are much bigger than the others, and therefore CS3 and BHL3 perform poorly according to RMSE. SHL2 performs quite well in comparison, with a cross-sectional correlation figure of 0.726, but reports a much lower error and standard deviation figure than CS3. Table 14 also suggests that the combination of SHL2 and BHL1 is the best choice while SHL2 provides a good alternative. Table 15 reports average cross-sectional correlations between true and estimated spreads generated by each of the estimators. All the estimators exhibit very high correlations. BHL3's correlation of 0.95 is the highest. CS3 and AR's correlation of 0.948 are the second strongest. SHL2 exhibits a moderate correlation of 0.864. Table 15 also reports currency-by-currency time series correlations between true and estimated spreads generated by each of the estimators. Similar to those reported in Corwin and Schultz (2012), all the estimators exhibit weaker correlations than in the cross-sectional case. BHL3, CS3 and SHL2 exhibit the highest correlations, which are 0.271, 0.265 and 0.198 respectively.

[Insert Tables 13 to 15 here]

Figure 2 illustrates an example of the estimates and the actual closing quoted spread in the form of the USD/JPY currency pair over a 50 month period. We can see that all estimators, except for SHL2, show negative estimates. CS2 and the combinations appear more volatile than SHL2, the combination of SHL2 and BHL1 has the lowest average estimated error.

[Insert figure 2 here]

4.4.1 Hotspot 12 currency pairs Dec 2015-Aug 2016

In this section, we evaluate the performance of the estimators using daily high and low prices and the effective time-weighted bid-ask spread data of 12 currency pairs sourced from Hotspot. Results are reported in Table 16. Hotspot is a large electronic communication network (ECN) platform for foreign exchange transactions conducted worldwide. We extract quotes and transaction data similar to that taken from the TAQ database. Trade volume weighted effective spreads are calculated for each pair over time the sample period begins in December 2015 and ends in August 2016. Spreads are arrived at through the matching of quote and transaction data. The trade volume weighted effective spread can be formally expressed as:

$$\begin{aligned} &2 \cdot (s_t - M_{t-1}) \quad \mathit{for\ buyer\ initiated\ trades} \\ &2 \cdot (M_{t-1} - s_t) \quad \mathit{for\ seller\ initiated\ trades} \end{aligned} \tag{42}$$

In order to reduce the possibility of errors in the data we eliminate outliers[f] and negative effective spreads. Table 16 displays the results of the pooled case where the 12 currencies over the entire sample period are examined. SHL2 is the best performing model in terms of RMSE, although its estimated error is very high (754%) but is similar to the others models. The standard deviation of SHL2 is the lowest (7.66) among

[f]Outliers are deemed to be those spreads which exceed the daily average by over 50 times.

the estimators tested, where each model either significantly over or under-estimates the spread. The currency-month pooled correlation coefficients for the estimated and true spreads for AR and BHL3 are 0.73 and 0.72 respectively; these are higher than the other models tested. However, their relative errors and standard deviations are greater in magnitude than the others. Therefore, AR and BHL3 are not the best performing estimators as the RMSEs place these at 13th and 14th in order of performance. The Roll estimator exhibits a high correlation coefficient (0.60), however, at least in half of the total cases, it obtains a negative number in the square root and we thus set it to zero. Although the combination of SHL2 and BHL1 is the second best in terms of RMSE, its correlation with the true spread is negative. SHL2's correlation coefficient is 0.24; this is acceptable in comparison with others. Also, its RMSE is the lowest amongst the models; therefore we rank this as the best performer. In Table 17, the average cross-sectional correlations between true and estimated spreads are reported across all currency pairs. AR and BHL3 exhibit the highest correlations, which are 0.8 and 0.77. SHL2's correlation is 0.636, performing slightly less well than AR and BHL3 in this instance but still at an adequate level. The average time series correlations between true and estimated spreads are reported across all currency pairs. BHL3 and AR exhibit the highest correlations, which are 0.51 and 0.48 respectively while with SHL2 the correlation is 0.04. The average time series correlations are much lower than those generated through cross-sectional analysis; this may be as a result of the time series being relatively short with its length being 12 months.

[Insert Tables 16 to 17]

4.4.2 Comparisons in equity markets TAQ data 2014

In this section, we use our chosen estimators to gauge spreads for the U.S equity market using the constituents of the S&P 1500 index as a sample. A snapshot of TAQ data, offers tick by tick pricing in 2014, which is used to calculate time-weighted quoted bid-ask spread and daily high and low prices. Tables 18 to 20 report the results of S&P 1500 (pooled case), S&P 600 (small cap), S&P 400 (mid cap), S&P 500 (large cap) stocks. In the pooled case, the SHL2 is the best performer in terms of RMSE results. It must be noted that all estimators significantly over or underestimate the spread.

SHL2 displays the smallest estimated error but underestimates the spread by 59% on average, while the next best performing model (CS3) has an error of 104%. In contrast to a relatively poor performance with FX simulations, CS3 is ranked second out of all the estimators. The combination of SHL2 and BHL1 takes the third place. The pooled equity-month correlation coefficients of BHL3 and SHL2 are 0.78 and 0.74 respectively; this is significantly higher than the others and suggests a high correlation with the true spread. CS3 slightly outperforms SHL2 in the case of equities listed on the S&P600 because of its -1% average error. However, it represents an isolated example because the error of the CS3 estimator for simulations, equity and FX sample tests throughout the paper is high. Table 22 reports the average ranking of the small, mid and large cap equity group cases. It is apparent that SHL2 is still the best choice for estimator while CS2 and the combination of SHL2 and BHL1 produce results that could offer a good alternative. Table 23 reports the equity-by-equity cross-sectional correlations for each of the estimators. BHL3 and SHL2's correlations are 0.82 and 0.75 respectively; these are significantly higher than the others reported through the testing. AR's correlation of 0.70 is the third strongest. Table 23 reports average time series correlations of the estimators. BHL3 and CS3's correlations are 0.34 and 0.32. SHL2's correlation is 0.09.

[Insert Tables18 to 23 here]

5 Application of SHL2

Moving beyond simulation, in this section, we demonstrate the application of SHL2 by using this estimator to gauge monthly average spreads for developed and emerging market stock exchanges. We find that the SHL2 acts as a good proxy for market liquidity as predictions of periods of intense uncertainty are often accompanied with low liquidity (high bid-ask spreads) levels in financial markets. In the samples we investigate, the data for the true spread is unavailable.

5.1 NYSE 1926-2015

Figure 3 shows the monthly average estimated spread of all US stock markets including New York Stock Exchange (NYSE), American Stock Exchange (AMEX) and teh

Nasdaq from 1926 to 2015; this is generated by SHL2 using daily CRSP data. The monthly average estimated spread of each market are also shown separately. The spread was relatively large in the years before 1935. A further period of low liquidity can be observed from 1970 to 1992, which is mainly caused by the low liquidity in the Nasdaq. In 2008 at the nadir of the global financial crisis, the SHL2 estimator recorded considerable lower liquidity levels through increased spreads, than in the years surrounding the event.

5.2 Non-US equity markets applications

Figures 4 to 6 show the monthly average estimated spreads for equities listed on the London, Hong Kong and Thai stock exchanges respectively. Data was obtained from Bloomberg.

We observe several increases in bid-ask spreads estimated by SHL2, i.e. transaction costs, around notable market events. For example, the average spread jumped to over 0.8% when the sterling crisis occurred in September 1992 (Figure 4). When the Asian financial crisis began in July 1997, transaction costs rose significantly in both the Thai and Hong Kong equity markets (Figures 5 and 6). The collapse of Lehman Brothers in September 2008 and the financial crisis which heralded drove a jump in spreads in almost all equity markets used in our samples. After the results of the Brexit referendum became clear in June 2016, transaction costs in the UK equity market also appeared to rise sharply.

[Insert figures 3 to 6 here]

The applications in this section suggest that SHL2 can act as an estimator that is sensitive enough to capture notable market events affecting transaction costs and as a consequence, liquidity levels. SHL2, as a spread estimator, can also be used as a liquidity measure in asset pricing models in a similar manner to that demonstrated in Corwin and Schultz (2012) and Abdi and Ranaldo (2017).

6 Conclusion

In this paper, we introduce two new low frequency bid-ask spread estimators which estimate the bid-ask spread using daily and two-day high and low prices. We show that using similar input data, our estimators, in particular, the sophisticated version, significantly outperforms both the latest and the popular models such as Abdi and Ranaldo (2017), Corwin and Schultz (2012) and Roll (1984) in terms of accuracy, efficiency, as well as cross-sectional and time series correlations.

We test the performance of estimators using comprehensive Monte Carlo simulation experiments under various 'signal to noise' ratios and different sampling frequencies. In addition, the cross-sectional properties of the estimators are also examined. Our estimators, BHL and SHL, appear to be unbiased throughout all tests carried out. By setting negative trials to zero which we label SHL2, we can obtain more efficient estimates; these are exhibited through lower standard errors. The results of simulation experiments suggest that our estimators outperform the AR, CS and Roll models. We demonstrate that SHL2 is the best single estimator in terms of accuracy and efficiency. We go further and test the performance of combinations of estimators against our own models and find that the AR, CS and Roll models also fail to match with ours in performance. The combinations of the estimators are useful as using these can address the problems associated with errors which often appear for individual models. The combination of basic and sophisticated high-low estimators (BHL1 and SHL2) perform well and offer a good alternative to using the single estimators, thereby avoiding the associated error risk.

We then move beyond simulation experiments to study the models using real world data for both foreign exchange and equity markets. By comparing the closing bid-ask spread of 22 currency pairs over 26 years, we find that our SHL2 model outperforms all the others including the AR, CS and Roll models in terms of the root mean square error (RMSE). We arrive at the same conclusion when we ran tests using trade and quote data of 12 currency pairs over 9 months and for equities listed on the S&P 1500 throughout 2014. In terms of correlation, BHL3, AR, CS3 and SHL2 all performed well as estimators.

In general, our BHL and SHL are the best spread estimators. Researchers can choose the estimator according to their needs: BHL3 is good for cases where the high correlation is the only requirement and SHL2 can be used for other cases especially when accuracy and efficiency is of particular importance.

In order to demonstrate the effectiveness of our model (SHL2) through applications, we provide an illustration of how this can be applied. We generate the average monthly bid-ask spreads for the US, UK, HK and Thai equity markets. We show how the estimated spreads follow a pattern that is in line with our expectations in that the transaction costs increase sharply during crises periods.

Similar to the CS model, our estimators also obtain the estimates of daily mid-price diffusion at the same time as when the spreads are estimated. Because the spread and the diffusion are estimated together, a good spread estimator is also a good diffusion estimator. Thus, our sophisticated version model (SHL2) also offers the best diffusion estimates. As our estimators are not designed for a particular market structure, further research could test and apply our suggested estimators to the bonds, futures and option markets. In particular, these may be interesting for the over-the-counter markets where quote data can be difficult to obtain.

7 Appendix

7.1 Proof of Proposition 3.1

When the components of the spread do not include feedback trading, inventory control or asymmetric information, we can consider that the spread and its estimates, and thus the estimated errors, are either serially independent or fixed. If an estimate of the spread $\widetilde{SP}_i \in A$ corresponds to $Var_i = max(B)$, it equals the true spread i.e. $\widetilde{SP}_i = SP$.

Proof The variance of two adjacent conjectures of mid-price returns is:

$$\begin{aligned} Var_i &= Var\left[\Delta\widetilde{M}_t\right] \\ &= E\left\{\left[\Delta\widetilde{M}_t - E\left(\Delta\widetilde{M}_t\right)\right]^2\right\} \end{aligned} \tag{43}$$

We will assume that the expectation of the value of the conjectural mid-prices is zero. Thus, the equation above can be rewritten as:

$$\begin{aligned} & Var_i = Var\left(\Delta\widetilde{M}_t\right) \\ & = E\left(\Delta\widetilde{M}_t^{\,2}\right) \\ & = E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)^2\right] \\ & = E\left[\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)\left(\Delta M_t + \tfrac{1}{2}\Omega BS_t - \tfrac{1}{2}\Omega BS_{t-1}\right)\right] \\ & = E\left(\Delta M_t^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t - \tfrac{1}{2}\Omega BS_{t-1}\Delta M_t\right) \\ & +E\left[\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{4}(\Omega BS_t)^2 - \tfrac{1}{4}\Omega^2 BS_t BS_{t-1}\right] \\ & -E\left[\tfrac{1}{2}\Delta M_t\Omega BS_{t-1} + \tfrac{1}{4}\Omega^2 BS_t BS_{t-1} - \tfrac{1}{4}(\Omega BS_{t-1})^2\right] \end{aligned} \tag{44}$$

Where Ω denotes the conjectural error which represents the difference between the conjectural mid-price and the true mid-price, alternately expressed as the difference between the conjectural spread and its true value. Formally, Ω is given as:

$$\Omega = \Delta\widetilde{M}_t - \Delta M_t = \widetilde{SP}_i - SP \tag{45}$$

The assumptions of this proposition imply that BS is independent of $\Delta\, M$ at all observation points, therefore many of the terms in (44) such as $E(\Delta\, M_t\; BS_t)$ and $E(\Delta\, M_{t-1}\; BS_t)$ equate to zero. Formally, we have:

$$
\begin{aligned}
&E(\Delta\, M_t\, BS_t) = 0 \\
&E(\Delta\, M_{t-1}\, BS_t) = 0 \\
&E(\Delta\, M_t\, BS_{t-1}) = 0 \\
&E\left(BS_t\, BS_{t-1}\right) = 0
\end{aligned} \tag{46}
$$

Furthermore, the variable BS is a binary variable (1 or -1), thus:

$$
E\left({BS_{t-1}}^2\right) = 1 \tag{47}
$$

Finally we obtain:

$$
\begin{aligned}
&Var_i = Var\left(\Delta\widetilde{M}_t\right) \\
&= E\left(\Delta {M_t}^2 + \tfrac{1}{2}\,\Omega^2\right)
\end{aligned} \tag{48}
$$

The final step of Equation (48) given above is the quadratic polynomial of the expectation of the error of the conjecture. For a given series, the first term $E(\Delta M_t^2)$ is a constant. We can surmise directly from this that when the error is zero (i.e. $\Omega = 0$), the second term $\frac{1}{2}\Omega^2$ is zero. Furthermore, when $\Omega = 0$, there is a global extreme for the right hand side polynomial in the final step, symmetrically, the left hand side of the equation $Var_i = Var\left(\Delta\widetilde{M}_t\right)$ is also at the extreme value. Formally this can be expressed as:

$$
\underset{\Omega}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = 0 \tag{49}
$$

When the conjectural error is zero, the conjectural spread becomes the true spread:

$$
\widetilde{SP}_i = SP + \Omega = SP \tag{50}
$$

Therefore the conjectural spread which maximises the covariance equals the true spread.

$$
\underset{\widetilde{SP}_i \in A}{arg\ max}\, Var\left(\Delta\widetilde{M}_t\right) = SP \tag{51}
$$

Q.E.D.

7.2 Proof of feedback bias

When feedback trading exists, we have:

$$
E(\Delta\, M_t\, BS_t) \neq 0 \tag{52}
$$

Substituting Equations (46), (47) and (52) and into Equation (44), we can obtain:

$$
\begin{aligned}
& Var_i = Var\left[\Delta\widetilde{M}_t\right] \\
& = E\left[(\Delta M_t)^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t - \tfrac{1}{2}\Omega BS_{t-1}\Delta M_t\right] \\
& +E\left[\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{4}(\Omega BS_t)^2 - \tfrac{1}{4}\Omega^2 BS_t BS_{t-1}\right] \\
& -E\left[\Delta M_t\tfrac{1}{2}\Omega BS_{t-1} + \tfrac{1}{4}\Omega^2 BS_t BS_{t-1} - \tfrac{1}{4}(\Omega BS_{t-1})^2\right] \\
& = E\left(\Delta M_t{}^2 + \tfrac{1}{2}\Omega BS_t\Delta M_t\right) \\
& +E\left(\tfrac{1}{2}\Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right) \\
& = E\left(\Delta M_t{}^2 + \Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right)
\end{aligned} \tag{53}
$$

Substituting $\Omega = \widetilde{SP} - SP = 0$ into the equation above, we have:

$$
\begin{aligned}
& Var_i = Var\left[\Delta\widetilde{M}_t\right] \\
& = E\left[(\Delta M_t)^2 + \Omega BS_t\Delta M_t + \tfrac{1}{2}\Omega^2\right] \\
& = E\left[(\Delta M_t)^2 + \left(\widetilde{SP} - SP\right) BS_t\Delta M_t + \tfrac{1}{2}\left(\widetilde{SP} - SP\right)^2\right]
\end{aligned} \tag{54}
$$

Using first order conditioning of Equation (54), we obtain:

$$SP = \widetilde{SP} - E\left(BS_t\,\Delta\,M_t\right) \tag{55}$$

Equation above suggests that when there is feedback trading, variance version of the BL estimator overestimates the spread.

7.3 Proof of Equation (31)

For each day, we choose at random either the daily high or low prices to calculate the daily price change. Thus, the probability of picking daily high (or low) price is 50% and there are four cases for the daily price changes with an equal likelihood which are as follows.

$$
\Delta\,M_{daily} = \begin{cases} H_t - H_{t-1} \;\; with \; \frac{1}{4} \; chance \\ H_t - L_{t-1} \;\; with \; \frac{1}{4} \; chance \\ L_t - H_{t-1} \;\; with \; \frac{1}{4} \; chance \\ L_t - L_{t-1} \;\; with \; \frac{1}{4} \; chance \end{cases} \tag{56}
$$

Thus $BS_{daily} \cdot \Delta\, M_{daily}$ is given as follows:

$$BS_{daily} \cdot \Delta\, M_{daily} = \begin{cases} BS_{daily} \cdot (H_t - H_{t-1}) \;\; with\; \frac{1}{4}\; chance \\ BS_{daily} \cdot (H_t - L_{t-1}) \;\; with\; \frac{1}{4}\; chance \\ BS_{daily} \cdot (L_t - H_{t-1}) \;\; with\; \frac{1}{4}\; chance \\ BS_{daily} \cdot (L_t - L_{t-1}) \;\; with\; \frac{1}{4}\; chance \end{cases} \tag{57}$$

When daily high (or low) price is picked, trading direction is known (Equation 4). Formally, we have:

$$BS_{daily} \cdot \Delta\, M_{daily} = \begin{cases} [1 \cdot (H_t - H_{t-1})] \;\; with\; \frac{1}{4}\; chance \\ [1 \cdot (H_t - L_{t-1})] \;\; with\; \frac{1}{4}\; chance \\ [-1 \cdot (L_t - H_{t-1})] \;\; with\; \frac{1}{4}\; chance \\ [-1 \cdot (L_t - L_{t-1})] \;\; with\; \frac{1}{4}\; chance \end{cases} \tag{58}$$

Taking the expectation of $BS_{daily} \cdot \Delta\, M_{daily}$, we obtain:

$$\begin{aligned} & E\left(BS_{daily} \cdot \Delta\, M_{daily}\right) = \tfrac{1}{4} \cdot E\left(H_t - H_{t-1}\right) + \tfrac{1}{4} \cdot E\left(H_t - L_{t-1}\right) \\ & -\tfrac{1}{4} \cdot E\left(L_t - H_{t-1}\right) - \tfrac{1}{4} \cdot E\left(L_t - L_{t-1}\right) \\ & = \tfrac{1}{2} E\left(H_t - L_t\right) \end{aligned} \tag{59}$$

7.4 A brief introduction to the AR, Roll and CS estimators

Researchers generally opt to use the Roll estimator and models[g] derived from it because they are easy to program. The Roll estimator is given by the following equation.

$$SP = 2\sqrt{-cov\left(\Delta s_t,\; \Delta s_{t-1}\right)} \tag{60}$$

According to Corwin and Schultz (2012), the CS estimator appears to be the best of low-frequency estimators including the Lesmond et al. (1999) estimator. Furthermore, our proposed model in this paper shares the same intuition with the CS estimator, therefore, the CS estimator is picked to examine. Squaring both sides of Equation (7), we have,

$$\begin{aligned} & \left(Range^{T}_{daily}\right)^2 = \left(Range^{M}_{daily} + SP\right)^2 \\ & = \left(Range^{M}_{daily}\right)^2 + 2\, Range^{M}_{daily} \cdot SP + (SP)^2 \end{aligned} \tag{61}$$

[g]Related models include Glosten and Harris (1988), Choi et al. (1988), Stoll (1989), George et al. (1991), Huang and Stoll (1997), Hasbrouck (2004, 2009) and Chen et al. (2016)

Similarly, squaring both sides of Equation (8), we have:

$$
\begin{aligned}
\left(Range^{T}_{twoday}\right)^2 &= \left(Range^{M}_{twoday} + SP\right)^2 \\
&= \left(Range^{M}_{twoday}\right)^2 + 2\,Range^{M}_{twoday} \cdot SP + (SP)^2
\end{aligned}
\tag{62}
$$

Corwin and Schultz (2012) assume that

$$
\begin{aligned}
E\left(Range^{T}_{twoday}\right)^2 &\approx E\left[\left(Range^{T}_{twoday}\right)^2\right] \\
E\left(Range^{T}_{daily}\right)^2 &\approx E\left[\left(Range^{T}_{daily}\right)^2\right]
\end{aligned}
\tag{63}
$$

One could solve the spread from the equation system and obtains:

$$
SP = \frac{2\left(e^{\alpha} - 1\right)}{1 + e^{\alpha}} \tag{64}
$$

where

$$
\alpha = \frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}} \tag{65}
$$

$$
\beta = E\left\{\sum_{J=0}^{1}\left(Range^{T}_{daily,t+J}\right)^2\right\};\ \gamma = \left(Range^{T}_{twoday}\right)^2 \tag{66}
$$

When the spread is small, $SP \approx \alpha$. We may use Equation (65) to estimate the spread.

Abdi and Ranaldo (2017) model incorporates the CS model into the Roll estimator. Formally, it can be expressed as follows:

$$
SP = 2\sqrt{\left(s_t - \eta_t\right)\left(s_t - \eta_{t+1}\right)} \tag{67}
$$

where η is the mid-point of the high and low prices. Formally, it is given by:

$$
\eta_t = \frac{H_t + L_t}{2} \tag{68}
$$

References

Abdi, F. and A. Ranaldo (2017). A Simple Estimation of Bid-Ask Spreads from Daily Close, High, and Low Prices. *University of St. Gallen, School of Finance Research Paper 2016/04*.

Bandi, F. M. and J. R. Russell (2006). Separating microstructure noise from volatility. *Journal of Financial Economics 79*(3), 655–692.

Banti, C., K. Phylaktis, and L. Sarno (2012). Global liquidity risk in the foreign exchange market. *Journal of International Money and Finance 31*(2), 267–291.

Bleaney, M. and Z. Li (2015). The performance of bid-ask spread estimators under less than ideal conditions Michael Bleaney and Zhiyong Li less than ideal conditions. *Studies in Economics and Finance 32*(1), 98–127.

Bleaney, M. and Z. Li (2016). A new spread estimator. *Review of Quantitative Finance and Accounting 47*(1), 179–211.

Chen, X., O. Linton, S. Schneeberger, and Y. Yi (2016). Simple Nonparametric Estimators for the Bid-Ask Spread in the Roll Model. *Cowles Foundation Discussion Paper No. 2033*, 1–59.

Choi, J. Y., D. Salandro, and K. Shastri (1988). On the Estimation of Bid-Ask Spreads: Theory and Evidence. *The Journal of Financial and Quantitative Analysis 23*(2), 219–230.

Chung, K. H. and H. Zhang (2014). A simple approximation of intraday spreads using daily data. *Journal of Financial Markets 17*(1), 94–120.

Corwin, S. A. and P. Schultz (2012). A Simple Way to Estimate Bid-Ask Spreads from Daily High and Low Prices. *Journal of Finance 67*(2), 719–760.

Fong, K. Y. L., C. W. Holden, and C. A. Trzcinka (2017). What Are The Best Liquidity Proxies For Global Research? *Review of Finance 6*, 1–22.

George, T. J., G. Kaul, and M. Nimalendran (1991). Estimation of the Bid-Ask Spread and its Components: A New Approach. *The Review of Financial Studies 4*(4), 623–656.

Glosten, L. R. and L. E. Harris (1988). Estimating the components of the bid/ask spread. *Journal of Financial Economics 21*(1), 123–142.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka (2009). Do liquidity measures measure liquidity? *Journal of Financial Economics 92*(2), 153–181.

Harris, L. (1990). Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator. *The Journal of Finance 45*(2), 579–590.

Hasbrouck, J. (2004). Liquidity in the Futures Pits: Inferring Market Dynamics from Incomplete Data. *The Journal of Financial and Quantitative Analysis 39*(2), 305–326.

Hasbrouck, J. (2009). Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data. *The Journal of Finance 64*(3), 1445–1477.

Holden, C. W. (2009). New low-frequency spread measures. *Journal of Financial Markets 12*(4), 778–813.

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

Huang, R. D. and H. R. Stoll (1997). The Components of the Bid-Ask Spread: A General Approach. *The Review of Financial Studies 10*(4), 995–1034.

Karnaukh, N., A. Ranaldo, and P. Soderlind (2015). Understanding FX Liquidity. *Review of Financial Studies 28*(11), 3073–3108.

Lesmond, D. A., J. P. Ogden, and C. A. Trzcinka (1999). A New Estimate of Transaction Costs. *The Review of Financial Studies 12*(5), pp. 1113–1141.

Mancini, L., A. Ranaldo, and J. Wrampelmeyer (2013). Liquidity in the Foreign Exchange Market: Measurement, Commonality, and Risk Premiums. *The Journal of Finance 68*(5), 1805–1841.

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *The Journal of Business 53*(1), 61–65.

Roll, R. (1984). A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market. *The Journal of Finance 39*(4), 1127–1139.

Stoll, H. R. (1989). Inferring the Components of the Bid-Ask Spread: Theory and Empirical Tests. *The Journal of Finance 44*(1), 115–134.

Table 1: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread= 1 (*0.001)	True spread/Midstd=0.00527					True spread/Midstd=0.0129				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking
SHL1	0.953	17.76%	41.177	41.176	8	0.899	-8.763%	6.513	6.514	7
SHL2∗ (negatives to be zero)	35.348	3391%	20.215	39.485	3	4.512	341%	3.609	4.968	1
BHL1⋆ (mean spreads)	0.756	-6.63%	40.958	40.958	6	0.906	-7.918%	6.495	6.496	5
BHL2⋆ (mean parameters)	0.793	0.97%	41.047	41.046	7	0.900	-8.702%	6.510	6.511	6
BHL3⋆ (negatives to be zero)	112.316	11122%	23.817	113.742	13	46.195	4519%	3.775	45.349	14
CS1ˆ (mean spreads)	21.221	2039%	37.204	42.426	9	9.155	817%	6.100	10.195	10
CS2ˆ (mean parameters)	2.643	185%	43.125	43.164	10	0.999	1.231%	7.139	7.139	8
CS3‡ (negatives to be zero)	82.901	8180%	19.591	84.112	12	34.266	3326%	3.221	33.415	12
ROLL	93.496	15269%	64.917	165.918	14	20.969	3758%	15.947	40.823	13
AR‡ (negatives to be zero)	74.679	7350%	22.047	76.738	11	30.438	2943%	3.503	29.635	11
Combination1 (BHL1+SHL2)/2	18.052	1693%	29.323	33.859	1	2.709	167%	4.972	5.244	2
Combination2 (SHL2+CS2)/2	18.996	1789%	30.546	35.397	2	2.756	171%	5.247	5.520	3
Combination3 (CS1+BHL1)/2	10.988	1016%	38.554	39.869	4	5.030	404%	6.262	7.454	9
Combination4 (BHL1+BHL2)/2	0.775	-2.55%	40.307	40.306	5	0.903	-8.297%	6.485	6.485	4

The standard divination of daily mid-price return is 0.1897. The true spread is fixed.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time intervals of daily and four-hours respectively. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Mean indicates the average of estimated spreads over 25000 replications. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation We also report the rankings of the estimators according to RMSE.

Table 2: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=2 (*0.001)	True spread/Midstd=0.0105					True spread/Midstd=0.0258				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking
SHL1	2.181	20.22%	20.577	20.578	8	1.909	-3.972%	3.272	3.272	7
SHL2∗ (negatives to be zero)	36.002	1680%	10.134	19.620	3	5.096	150%	1.951	2.459	1
BHL1⋆ (mean spreads)	1.985	8.91%	20.431	20.431	6	1.909	-4.022%	3.263	3.263	5
BHL2⋆ (mean parameters)	2.169	19.27%	20.546	20.547	7	1.911	-3.887%	3.270	3.270	6
BHL3⋆ (negatives to be zero)	113.081	5550%	11.873	56.758	13	46.773	2238%	1.912	22.466	14
CS1ˆ (mean spreads)	22.457	1032%	18.541	21.217	9	10.116	406%	3.069	5.092	10
CS2ˆ (mean parameters)	3.767	97.52%	21.513	21.535	10	1.967	-1.012%	3.593	3.593	8
CS3‡ (negatives to be zero)	83.778	4085%	9.757	41.997	12	34.880	1644%	1.635	16.518	12
ROLL	93.022	7581%	32.519	82.487	14	21.411	1839%	8.006	20.059	13
AR‡ (negatives to be zero)	74.919	3637%	10.969	37.985	11	30.540	1426%	1.756	14.372	11
Combination1 (BHL1+SHL2)/2	18.993	845%	14.672	16.929	1	3.503	72.871%	2.571	2.672	2
Combination2 (SHL2+CS2)/2	19.884	889%	15.277	17.674	2	3.531	74.460%	2.712	2.812	3
Combination3 (CS1+BHL1)/2	12.221	520%	19.240	19.930	4	6.012	201%	3.148	3.736	9
Combination4 (BHL1+BHL2)/2	2.077	14.08%	20.155	20.155	5	1.910	-3.949%	3.257	3.258	4

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (1).

Table 3: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=6 (*0.001)	True spread/Midstd=0.0316					True spread/Midstd=0.0775				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking
SHL1	5.896	1.82%	6.900	6.900	8	5.818	-2.781%	1.080	1.081	7
SHL2∗ (negatives to be zero)	38.141	529%	3.523	6.354	3	7.634	25.770%	0.792	0.833	1
BHL1⋆ (mean spreads)	5.805	-0.41%	6.831	6.831	6	5.821	-2.709%	1.080	1.080	5
BHL2⋆ (mean parameters)	5.925	1.97%	6.896	6.895	7	5.817	-2.782%	1.080	1.080	6
BHL3⋆ (negatives to be zero)	115.153	1818%	3.966	18.608	13	48.999	717%	0.646	7.195	14
CS1ˆ (mean spreads)	25.936	335%	6.219	7.063	9	13.797	130%	1.014	1.650	10
CS2ˆ (mean parameters)	7.377	25.97%	7.223	7.227	10	5.714	-4.482%	1.184	1.184	8
CS3‡ (negatives to be zero)	85.942	1331%	3.315	13.713	12	37.168	519%	0.561	5.224	12
ROLL	93.559	2455%	10.843	26.840	14	21.573	546%	2.646	6.071	13
AR‡ (negatives to be zero)	74.826	1144%	3.659	12.013	11	30.687	411%	0.586	4.154	11
Combination1 (BHL1+SHL2)/2	21.973	264%	4.975	5.633	1	6.728	11.535%	0.927	0.935	2
Combination2 (SHL2+CS2)/2	22.759	277%	5.193	5.888	2	6.674	10.668%	0.971	0.977	3
Combination3 (CS1+BHL1)/2	15.870	167%	6.443	6.656	4	9.809	63.730%	1.041	1.221	9
Combination4 (BHL1+BHL2)/2	5.865	0.82%	6.750	6.750	5	5.819	-2.740%	1.076	1.077	4

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (1).

Table 4: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=10 (*0.001)	True spread/Midstd=0.0527					True spread/Midstd=0.129				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking
SHL1	9.884	0.82%	4.143	4.143	8	9.728	-2.641%	0.657	0.657	6
SHL2∗ (negatives to be zero)	40.456	301%	2.182	3.714	3	10.699	6.274%	0.552	0.556	1
BHL1⋆ (mean spreads)	9.829	-0.04%	4.097	4.097	6	9.730	-2.629%	0.656	0.657	5
BHL2⋆ (mean parameters)	9.833	0.15%	4.133	4.133	7	9.729	-2.626%	0.657	0.658	7
BHL3⋆ (negatives to be zero)	117.638	1076%	2.385	11.017	13	51.223	412%	0.393	4.141	14
CS1ˆ (mean spreads)	29.716	199%	3.725	4.222	9	17.474	74.802%	0.617	0.970	10
CS2ˆ (mean parameters)	11.212	14.09%	4.313	4.315	10	9.454	-5.372%	0.718	0.720	8
CS3‡ (negatives to be zero)	88.401	783%	2.014	8.085	12	39.539	295%	0.349	2.974	12
ROLL	93.914	1433%	6.524	15.746	14	21.875	293%	1.611	3.343	13
AR‡ (negatives to be zero)	75.142	650%	2.200	6.860	11	30.997	210%	0.355	2.128	11
Combination1 (BHL1+SHL2)/2	25.142	150%	3.018	3.371	1	10.214	1.822%	0.601	0.601	2
Combination2 (SHL2+CS2)/2	25.834	157%	3.140	3.512	2	10.076	0.455%	0.626	0.626	3
Combination3 (CS1+BHL1)/2	19.772	99.33%	3.862	3.987	4	13.602	36.083%	0.633	0.729	9
Combination4 (BHL1+BHL2)/2	9.831	0.06%	4.048	4.048	5	9.729	-2.628%	0.655	0.655	4

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (1).

Table 5: Simulation experiments: Comparison of the estimates over 25000 months

	Daily (MidStd= 189.7*0.001) 20 observations per month					Four hours (MidStd =77.5*0.001) 120 observations per month				
True spread=30 (*0.001)	True spread/Midstd=0.158					True spread/Midstd=0.387				
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking
SHL1	29.607	-0.61%	1.387	1.387	8	29.809	-0.603%	0.218	0.218	5
SHL2∗	52.870	74.94%	0.833	1.121	1	29.818	-0.588%	0.218	0.218	4
(negatives to be zero)										
BHL1⋆	29.567	-0.92%	1.375	1.375	6	29.820	-0.573%	0.218	0.218	3
(mean spreads)										
BHL2⋆	29.595	-0.74%	1.385	1.385	7	29.809	-0.607%	0.218	0.218	6
(mean parameters)										
BHL3⋆	128.947	330%	0.826	3.398	13	63.517	112%	0.142	1.126	14
(negatives to be zero)										
CS1ˆ	48.244	61.34%	1.249	1.391	9	36.619	22.091%	0.206	0.302	11
(mean spreads)										
CS2ˆ	30.054	0.83%	1.439	1.439	10	29.055	-3.116%	0.237	0.239	10
(mean parameters)										
CS3‡	100.202	234%	0.724	2.447	12	52.595	75.295%	0.135	0.765	13
(negatives to be zero)										
ROLL	95.550	416%	2.206	4.706	14	30.286	47.501%	0.582	0.751	12
AR‡	76.747	155%	0.750	1.724	11	35.313	17.669%	0.131	0.220	7
(negatives to be zero)										
Combination1	41.218	36.99%	1.067	1.129	2	29.819	-0.581%	0.217	0.217	1
(BHL1+SHL2)/2										
Combination2	41.462	37.88%	1.103	1.166	3	29.437	-1.856%	0.226	0.226	8
(SHL2+CS2)/2										
Combination3	38.905	30.19%	1.296	1.331	4	33.219	10.758%	0.211	0.237	9
(CS1+BHL1)/2										
Combination4	29.581	-0.83%	1.358	1.358	5	29.814	-0.590%	0.218	0.218	2
(BHL1+BHL2)/2										

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (1).

Table 6: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=11.02 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0581					
Range from 2.00 to 20.00 (*0.001)	75000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	10.461	-2.53%	5.612	5.612	7	0.113
SHL2∗ (negatives to be zero)	40.867	378%	4.006	5.506	5	**0.127**
BHL1⋆ (mean spreads)	10.429	-3.79%	5.560	5.560	6	0.115
BHL2⋆ (mean parameters)	10.438	-2.87%	5.618	5.618	8	0.114
BHL3⋆ (negatives to be zero)	117.774	1343%	10.005	16.748	13	0.112
CS1ˆ (mean spreads)	30.265	242%	5.339	5.863	10	0.119
CS2ˆ (mean parameters)	11.820	14.74%	5.849	5.851	9	0.107
CS3‡ (negatives to be zero)	88.586	979.10%	7.484	12.324	12	**0.136**
ROLL	93.103	1759%	14.736	22.950	14	0.003
AR‡ (negatives to be zero)	74.796	825.24%	6.948	10.788	11	0.008
Combination1 (BHL1+SHL2)/2	25.648	186%	4.286	4.673	1	0.124
Combination2 (SHL2+CS2)/2	26.343	195%	4.462	4.871	2	0.117
Combination3 (CS1+BHL1)/2	20.347	119%	5.314	5.445	3	0.118
Combination4 (BHL1+BHL2)/2	10.434	-3.38%	5.499	5.499	4	0.116

The standard divination of daily mid-price return is 0.1897

Mean indicates the average of estimated spreads over 75000 months. The true spread changes every month ranging from 0.002 to 0.02.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The column 'Ranking' reports the rankings of the estimators according to values produced in column RMSE.

This table reports the results of the time interval of daily. *Midstd* represents the standard deviation of mid-price returns over the relevant interval. For each time interval, there are five panels which report the summary statistics and the results of the estimators respectively. Outliers of relative errors, the highest and lowest 1% of the relative estimated errors, are trimmed off before further calculation.

Table 7: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=3.57 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0188					
Range from 2.00 to 5.13 (*0.001)	15000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	3.403	3.51%	12.511	12.511	7	0.010
SHL2∗	36.725	983%	6.814	11.957	3	0.010
(negatives to be zero)						
BHL1⋆	3.253	-3.09%	12.375	12.375	6	0.013
(mean spreads)						
BHL2⋆	3.494	4.31%	12.543	12.543	8	0.015
(mean parameters)						
BHL3⋆	113.768	3299.10%	11.395	34.903	13	**0.019**
(negatives to be zero)						
CS1ˆ	23.663	605.75%	11.388	12.899	9	0.017
(mean spreads)						
CS2ˆ	4.996	47.54%	13.054	13.062	10	**0.018**
(mean parameters)						
CS3‡	84.477	2423%	8.860	25.801	12	0.012
(negatives to be zero)						
ROLL	93.085	4448%	22.687	49.929	14	0.003
AR‡	74.800	2133%	8.872	23.101	11	-0.005
(negatives to be zero)						
Combination1	19.989	489%	9.053	10.291	1	0.013
(BHL1+SHL2)/2						
Combination2	20.861	515%	9.429	10.742	2	0.016
(SHL2+CS2)/2						
Combination3	13.458	301%	11.701	12.082	4	0.015
(CS1+BHL1)/2						
Combination4	3.373	0.36%	12.252	12.251	5	0.015
(BHL1+BHL2)/2						

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.002 to 0.00513.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (6).

Table 8: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=6.71 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0354					
Range from 5.13 to 8.29 (*0.001)	15000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	6.533	0.62%	6.311	6.310	8	0.022
SHL2∗	38.490	476%	3.296	5.786	3	0.020
(negatives to be zero)						
BHL1⋆	6.521	-1.06%	6.268	6.268	6	**0.026**
(mean spreads)						
BHL2⋆	6.458	-0.96%	6.294	6.294	7	0.024
(mean parameters)						
BHL3⋆	115.408	1646%	4.232	16.993	13	0.024
(negatives to zero)						
CS1ˆ	26.490	302%	5.709	6.457	9	0.025
(mean spreads)						
CS2ˆ	7.966	21.36%	6.610	6.613	10	0.023
(mean parameters)						
CS3‡	86.192	1203%	3.458	12.518	12	0.022
(negatives to zero)						
ROLL	93.129	2224%	10.244	24.486	14	0.001
AR‡	74.634	1028%	3.641	10.905	11	0.005
(negatives to be zero)						
Combination1	22.505	237%	4.573	5.152	1	0.025
(BHL1+SHL2)/2						
Combination2	23.228	248%	4.765	5.374	2	0.022
(SHL2+CS2)/2						
Combination3	16.505	150%	5.908	6.095	4	**0.026**
(CS1+BHL1)/2						
Combination4	6.489	-0.96%	6.177	6.177	5	**0.026**
(BHL1+BHL2)/2						

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.00513 to 0.00829.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (6).

Table 9: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=9.86 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0520					
Range from 8.29 to 11.44 (*0.001)	15000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	8.527	-11.13%	4.249	4.250	7	0.019
SHL2∗ (negatives to be zero)	39.773	302%	2.218	3.744	3	**0.021**
BHL1⋆ (mean spreads)	8.584	-10.81%	4.222	4.223	6	0.018
BHL2⋆ (mean parameters)	8.323	-13.21%	4.251	4.253	8	0.018
BHL3⋆ (negatives to be zero)	116.896	1093%	2.644	11.244	13	0.020
CS1ˆ (mean spreads)	28.435	192%	3.836	4.289	9	0.018
CS2ˆ (mean parameters)	9.763	1.38%	4.445	4.445	10	0.016
CS3‡ (negatives to be zero)	87.575	793%	2.191	8.230	12	**0.021**
ROLL	92.115	1447%	6.774	15.979	14	-0.002
AR‡ (negatives to be zero)	74.396	658%	2.345	6.987	11	0.004
Combination1 (BHL1+SHL2)/2	24.178	145%	3.090	3.415	1	0.020
Combination2 (SHL2+CS2)/2	24.768	152%	3.214	3.554	2	0.018
Combination3 (CS1+BHL1)/2	18.509	90.50%	3.978	4.079	4	0.018
Combination4 (BHL1+BHL2)/2	8.453	-12.01%	4.169	4.171	5	0.018

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.00829 to 0.0114.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (6).

Table 10: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=12.99 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0685					
Range from 11.45 to 14.55 (*0.001)	15000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	12.128	-4.66%	3.218	3.218	8	0.020
SHL2$\ast$	41.823	220%	1.730	2.798	3	0.025
(negatives to be zero)						
BHL1$\star$	12.094	-5.31%	3.198	3.198	6	0.018
(mean spreads)						
BHL2$\star$	12.081	-5.39%	3.217	3.217	7	0.020
(mean parameters)						
BHL3$\star$	118.783	818%	1.939	8.403	13	**0.028**
(negatives to be zero)						
CS1^	31.829	147%	2.907	3.258	9	0.020
(mean spreads)						
CS2^	13.532	5.84%	3.363	3.364	10	0.018
(mean parameters)						
CS3‡	89.523	591%	1.642	6.137	12	**0.027**
(negatives to be zero)						
ROLL	92.732	1086%	5.084	11.992	14	0.010
AR‡	74.802	477%	1.745	5.080	11	0.012
(negatives to be zero)						
Combination1	26.958	107%	2.370	2.601	1	0.022
(BHL1+SHL2)/2						
Combination2	27.678	113%	2.463	2.709	2	0.021
(SHL2+CS2)/2						
Combination3	21.961	70.83%	3.013	3.095	4	0.020
(CS1+BHL1)/2						
Combination4	12.087	-5.39%	3.156	3.156	5	0.020
(BHL1+BHL2)/2						

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.0114 to 0.0146.

$\ast$In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

$\star$The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

^The monthly CS estimates can be calculated using the two methods described in note $\star$.

‡ Estimates are calculated in a manner similar to that described in notes $\ast$.

The other settings are the same as Table (6).

Table 11: Simulation experiments: Cross-sectional properties of the estimates

Mean true spread=16.10 (*0.001)	Daily (MidStd= 189.7*0.001) Truespread/Midstd=0.0849					
Range from 14.55 to 17.65 (*0.001)	15000 months 20 observations per month					
	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	15.941	0.68%	2.590	2.590	7	0.010
SHL2∗	44.044	172%	1.425	2.232	3	0.008
(negatives to be zero)						
BHL1⋆	15.980	0.71%	2.577	2.577	6	**0.013**
(mean spreads)						
BHL2⋆	15.981	0.82%	2.596	2.596	8	0.011
(mean parameters)						
BHL3⋆	120.740	652%	1.559	6.699	13	**0.012**
(negatives to be zero)						
CS1^	35.399	122%	2.346	2.642	9	0.009
(mean spreads)						
CS2^	17.234	8.60%	2.718	2.720	10	0.009
(mean parameters)						
CS3‡	91.672	470%	1.316	4.885	12	0.011
(negatives to be zero)						
ROLL	93.978	862%	4.100	9.543	14	-0.010
AR‡	74.862	365%	1.409	3.915	11	-0.010
(negatives to be zero)						
Combination1	30.012	86.30%	1.928	2.112	1	**0.012**
(BHL1+SHL2)/2						
Combination2	30.639	90.25%	2.008	2.201	2	0.009
(SHL2+CS2)/2						
Combination3	25.689	61.12%	2.431	2.507	4	0.011
(CS1+BHL1)/2						
Combination4	15.981	0.78%	2.546	2.546	5	**0.012**
(BHL1+BHL2)/2						

Mean indicates the average of estimated spreads over 15000 months. The true spread changes every month ranging from 0.0146 to 0.0177.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

^The monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

The other settings are the same as Table (6).

Table 12: Simulation experiments: Average Ranking according to RMSE

	All cases according to simulations in sections 4.2.1 and 4.2.2	Fixed spread cases according to simulations in section 4.2.1	Cross-sectional cases according to simulations in section 4.2.2
SHL1	7.3	7.2	7.4
SHL2$*$ (negatives to be zero)	2.4	2.1	3
BHL1 (mean spreads) $\star$	5.5	5.3	6
BHL2 (mean parameters) $\star$	6.9	6.6	7.6
BHL3 (negatives to be zero) $\star$	13.3	13.5	13
CS1ˆ(mean spreads)	9.4	9.6	9
CS2ˆ(mean parameters)	9.5	9.2	10
CS3‡ (negatives to be zero)	12.1	12.1	12
ROLL	13.6	13.4	14
AR‡ (negatives to be zero)	10.7	10.6	11
Combination1 (BHL1+SHL2)/2	1.3	1.5	1
Combination2 (SHL2+CS2)/2	2.7	3.1	2
Combination3 (CS1+BHL1)/2	5.7	6.5	4
Combination4 (BHL1+BHL2)/2	4.5	4.3	5

This table reports the average ranking of the estimators and the combinations in simulations experiments according to the columns of ranking in Tables 1 to 11.

Table 13: DataStream 22 Currency pairs from 1990 to 2016 pooled

Mean closing spread = 14.24 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	7.392	-44.49%	2.356	2.397	9	0.043
SHL2∗ (negatives to be zero)	29.880	90.49%	1.398	1.665	2	0.726
BHL1⋆ (mean spreads)	9.044	-37.01%	2.243	2.274	6	0.076
BHL2⋆ (mean parameters)	8.636	-38.69%	2.368	2.399	10	0.069
BHL3⋆ (negatives to be zero)	94.082	505%	3.835	6.343	13	**0.869**
CS1ˆ (mean spreads)	29.406	91.16%	1.993	2.192	5	0.628
CS2ˆ (mean parameters)	16.944	17.05%	2.351	2.357	8	0.251
CS3‡ (negatives to be zero)	60.026	313%	2.488	4.000	12	0.876
ROLL	64.439	580%	5.548	8.031	14	0.591
AR‡ (negatives to be zero)	54.120	253%	2.441	3.517	11	**0.832**
Combination1 (BHL1+SHL2)/2	19.462	26.73%	1.637	1.659	1	0.423
Combination2 (SHL2+CS2)/2	23.412	53.77%	1.744	1.825	3	0.505
Combination3 (CS1+BHL1)/2	19.225	26.71%	2.033	2.050	4	0.367
Combination4 (BHL1+BHL2)/2	8.838	-37.79%	2.249	2.280	7	0.074

The results above refer to the testing of the following currency pairs: AUD/USD, CAD/USD, CHF/USD, DKK/USD, EUR/USD, GBP/USD, NOK/USD, SEK/USD, USD/JPY, USD/NZD, AUD/EUR, CAD/GBP, CHF/EUR, DKK/GBP, EUR/GBP, EUR/JPY, GBP/AUD, GBP/JPY, NOK/EUR, NOK/GBP, SEK/EUR, SEK/GBP. Daily closing spreads are used to calculate the monthly benchmark bid-ask spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 14: Currency-by-currency Average Ranking DataStream from 1990 to 2016

	All cases
SHL1	8
SHL2	1.7
(negatives to be zero)∗	
BHL1	6
(mean spreads)⋆	
BHL2	8.3
(mean parameters)⋆	
BHL3	13.1
(negatives to be zero)⋆	
CS1	7.3
(mean spreads)ˆ	
CS2	8.2
(mean parameters)ˆ	
CS3‡	12
(negatives to be zero)	
ROLL	13.9
AR‡	10
(negatives to be zero)	
Combination1	1.5
(SHL2+BHL1)/2	
Combination2	3.8
(SHL2+CS2)/2	
Combination3	5.5
(CS1+BHL1)/2	
Combination4	5.6
(BHL1+BHL2)/2	

This table reports the average ranking of the estimators and the combinations for 22 currency pairs according to RMSE.
Currency pairs used in this table are listed in Table 13

Table 15: Currency-by-currency average correlation DataStream from 1990 to 2016

	Time series correlation	Cross-sectional correlation
SHL1	0.091	0.147
SHL2∗ (negatives to be zero)	0.198	0.864
BHL1⋆ (mean spreads)	0.074	0.176
BHL2⋆ (mean parameters)	0.092	0.179
BHL3⋆ (negatives to be zero)	**0.271**	**0.951**
CS1ˆ(mean spreads)	0.158	0.636
CS2ˆ(mean parameters)	0.125	0.297
CS3‡ (negatives to be zero)	**0.265**	**0.948**
ROLL	0.108	0.543
AR‡(negatives to be zero)	0.254	**0.948**
Combination1(BHL1+SHL2)/2	0.127	0.443
Combination2(SHL2+CS2)/2	0.158	0.518
Combination3(CS1+BHL1)/2	0.118	0.4
Combination4(BHL1+BHL2)/2	0.085	0.187

This table reports the average time-series and cross-sectional correlation of the estimators and the combinations.

Currency pairs used in this table are listed in Table 13

Daily closing spreads are used to calculate the monthly benchmark bid-ask spread.

Highest two correlation coefficients are made bold.

Table 16: Hotspot 12 Currency pairs from 2015.12 to 2016.8

Effective Spread = 3.012 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	-41.563	-1632%	24.020	28.944	10	-0.477
SHL2∗ (negatives to be zero)	34.085	754%	7.659	10.718	1	0.241
BHL1⋆ (mean spreads)	-31.939	-1204%	21.287	24.364	7	-0.500
BHL2⋆ (mean parameters)	-38.304	-1618%	25.328	29.955	11	-0.436
BHL3⋆ (negatives to be zero)	169.091	5768%	33.146	66.443	14	**0.716**
CS1ˆ (mean spreads)	26.925	876%	10.814	13.874	5	0.088
CS2ˆ (mean parameters)	12.673	492%	14.244	15.004	6	-0.086
CS3‡ (negatives to be zero)	74.603	2515%	12.350	27.990	9	0.567
ROLL	99.593	4559%	28.853	53.834	12	0.596
AR‡ (negatives to be zero)	159.270	5191%	25.001	57.561	13	**0.732**
Combination1 (BHL1+SHL2)/2	1.073	-229%	11.608	11.777	3	-0.218
Combination2 (SHL2+CS2)/2	23.379	627%	9.579	11.412	2	0.078
Combination3 (CS1+BHL1)/2	-2.507	-163%	13.433	13.467	4	-0.296
Combination4 (BHL1+BHL2)/2	-35.121	-1411%	23.044	26.924	8	-0.476

The results above refer to the testing of the following currency pairs: AUD/USD, EUR/GBP, EUR/JPY, EUR/SEK, EUR/USD, GBP/USD, NZD/USD, USD/CAD, USD/CHF, USD/CHF, USD/JPY, USD/MXN, USD/ZAR. Tick by tick transaction and quoted data are used to generate the monthly effective spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 17: Currency-by-currency average correlation Hotspot from 2015.12 to 2016.8

	Time series correlation	Cross-sectional correlation
SHL1	-0.409	-0.325
SHL2 (negatives to be zero) $*$	0.035	0.636
BHL1 (mean spreads) $\star$	-0.292	-0.501
BHL2 (mean parameters) $\star$	-0.298	-0.331
BHL3 (negatives to be zero) $\star$	**0.514**	**0.772**
CS1 (mean spreads) ˆ	0.058	0.338
CS2 (mean parameters) ˆ	0.042	-0.133
CS3‡ (negatives to be zero)	0.321	0.739
ROLL	0.239	0.575
AR‡ (negatives to be zero)	**0.480**	**0.800**
Combination1 (BHL1+SHL2)/2	-0.235	-0.156
Combination2 (SHL2+CS2)/2	0.081	0.192
Combination3 (CS1+BHL1)/2	-0.194	-0.305
Combination4 (BHL1+BHL2)/2	-0.297	-0.383

This table reports the average time-series and cross-sectional correlations of the estimators and the combinations.

Currency pairs used in this table are listed in Table 16

Tick by tick transaction and quoted data are used to generate the monthly effective spread.

Highest two correlation coefficients are made bold.

Table 18: TAQ (Quoted Spread) S&P 1500 2014.01-2014.12

Quoted Spread = 310.2 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	-264	-356%	3.758	5.177	11	-0.086
SHL2∗ (negatives to be zero)	70	-59%	0.512	0.779	1	**0.743**
BHL1⋆ (mean spreads)	-277	-367%	3.764	5.259	12	-0.083
BHL2⋆ (mean parameters)	-256	-351%	3.751	5.135	9	-0.082
BHL3⋆ (negatives to be zero)	362	193%	2.648	3.274	7	**0.779**
CS1ˆ (mean spreads)	-44	-211%	2.277	3.106	5	0.301
CS2ˆ (mean parameters)	-236	-375%	4.332	5.732	13	0.009
CS3‡ (negatives to be zero)	222	104%	1.849	2.119	2	0.527
ROLL	426	466%	6.109	7.683	14	0.276
AR‡ (negatives to be zero)	306	174%	2.748	3.255	6	0.679
Combination1 (BHL1+SHL2)/2	-104	-212%	1.903	2.852	3	0.030
Combination2 (SHL2+CS2)/2	-83	-217%	2.208	3.094	4	0.303
Combination3 (CS1+BHL1)/2	-161	-290%	2.985	4.164	8	-0.039
Combination4 (BHL1+BHL2)/2	-267	-359%	3.728	5.175	10	-0.083

Tick by tick quoted data are used to generate the monthly quoted spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 19: TAQ (Quoted Spread) S&P 500 2014.01-2014.12

Quoted Spread = 202.3 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	-462	-632%	5.317	8.256	11	-0.273
SHL2∗ (negatives to be zero)	55	-40%	0.743	0.843	1	0.527
BHL1⋆ (mean spreads)	-468	-646%	5.264	8.329	12	-0.292
BHL2⋆ (mean parameters)	-451	-624%	5.358	8.227	9	-0.286
BHL3⋆ (negatives to be zero)	416	415%	3.253	5.270	6	**0.858**
CS1ˆ (mean spreads)	-106	-348%	3.322	4.808	5	0.204
CS2ˆ (mean parameters)	-370	-667%	6.230	9.127	13	-0.275
CS3‡ (negatives to be zero)	241	256%	2.287	3.431	2	0.623
ROLL	569	981%	8.223	12.797	14	0.368
AR‡ (negatives to be zero)	402	413%	3.438	5.375	7	**0.814**
Combination1 (BHL1+SHL2)/2	-207	-342%	2.727	4.373	3	-0.256
Combination2 (SHL2+CS2)/2	-157	-353%	3.236	4.788	4	-0.097
Combination3 (CS1+BHL1)/2	-287	-499%	4.218	6.532	8	-0.269
Combination4 (BHL1+BHL2)/2	-460	-635%	5.261	8.244	10	-0.290

Tick by tick quoted data are used to generate the monthly quoted spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 20: TAQ (Quoted Spread) S&P 400 2014.01-2014.12

Quoted Spread = 365 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	-225	-304%	2.892	4.199	10	-0.011
SHL2∗ (negatives to be zero)	78	-61%	0.454	0.757	1	**0.829**
BHL1⋆ (mean spreads)	-240	-317%	2.911	4.302	12	0.025
BHL2⋆ (mean parameters)	-213	-298%	2.905	4.165	9	0.019
BHL3⋆ (negatives to be zero)	373	155%	1.987	2.519	5	0.817
CS1ˆ (mean spreads)	-45	-189%	1.828	2.626	7	0.369
CS2ˆ (mean parameters)	-225	-321%	3.445	4.710	13	0.101
CS3‡ (negatives to be zero)	219	79%	1.398	1.603	2	0.584
ROLL	393	372%	4.561	5.885	14	0.459
AR‡ (negatives to be zero)	295	129%	1.984	2.367	3	**0.867**
Combination1 (BHL1+SHL2)/2	-81	-188%	1.497	2.406	4	0.292
Combination2 (SHL2+CS2)/2	-73	-191%	1.782	2.611	6	0.525
Combination3 (CS1+BHL1)/2	-142	-253%	2.328	3.441	8	0.093
Combination4 (BHL1+BHL2)/2	-226	-308%	2.884	4.215	11	0.022

Tick by tick quoted data are used to generate the monthly quoted spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 21: TAQ (Quoted Spread) S&P 600 2014.01-2014.12

Quoted Spread = 364 (*0.001)	Mean*0.001	Rel-Err-Mean	Rel-Err-Std	RMSE	Ranking	Correlation
SHL1	-123	-176%	1.324	2.203	10	-0.024
SHL2∗ (negatives to be zero)	76	-72%	0.290	0.776	2	0.614
BHL1⋆ (mean spreads)	-139	-184%	1.351	2.282	12	-0.040
BHL2⋆ (mean parameters)	-120	-174%	1.315	2.181	9	-0.042
BHL3⋆ (negatives to be zero)	307	40%	1.033	1.107	4	**0.745**
CS1ˆ (mean spreads)	9	-120%	0.837	1.462	6	0.352
CS2ˆ (mean parameters)	-129	-184%	1.644	2.468	13	0.056
CS3‡ (negatives to be zero)	207	-1%	0.733	0.733	1	0.576
ROLL	324	140%	2.236	2.639	14	0.308
AR‡ (negatives to be zero)	232	14%	0.975	0.985	3	**0.691**
Combination1 (BHL1+SHL2)/2	-31	-128%	0.712	1.462	5	0.150
Combination2 (SHL2+CS2)/2	-27	-128%	0.869	1.547	7	0.231
Combination3 (CS1+BHL1)/2	-65	-152%	1.068	1.859	8	0.093
Combination4 (BHL1+BHL2)/2	-129	-179%	1.320	2.223	11	-0.041

Tick by tick quoted data are used to generate the monthly quoted spread.

∗In the instance where the SHL estimate in a trail is a negative value, we set all negative estimated spreads in a trail to zero.

⋆The BHL estimates can be calculated using two methods: (1) calculate the two-day interval spread for one equity finding the monthly mean for the spread (reported as 'BHL1 mean spreads' in the table above); 2) calculate the average daily and two day interval range each month and then calculate the spread (reported as 'BHL2 mean parameters' above).

ˆThe monthly CS estimates can be calculated using the two methods described in note ⋆.

‡ Estimates are calculated in a manner similar to that described in notes ∗.

Table 22: Average Ranking S&P 2014.01-2014.12

	All cases
SHL1	10.3
SHL2 (negatives to be zero)$*$	1.3
BHL1 (mean spreads)$\star$	12
BHL2 (mean parameters)$\star$	9
BHL3 (negatives to be zero)$\star$	5
CS1 (mean spreads)ˆ	6
CS2 (mean parameters)ˆ	13
CS3‡ (negatives to be zero)	1.7
ROLL	14
AR‡ (negatives to be zero)	4.3
Combination1 (SHL2+BHL1)/2	4.0
Combination2 (SHL2+CS2)/2	5.7
Combination3 (CS1+BHL1)/2	8.0
Combination4 (BHL1+BHL2)/2	10.7

This table reports the average ranking of the estimators and the combinations according to the columns of ranking in Tables 19 to 21.

Table 23: Stock-by-stock average correlation S&P 1500 2014.01-2014.12

	Time series correlation	Cross-sectional correlation
SHL1	-0.052	-0.105
SHL2 (negatives to be zero) $*$	0.088	**0.749**
BHL1 (mean spreads) $\star$	-0.039	-0.111
BHL2 (mean parameters) $\star$	-0.053	-0.106
BHL3 (negatives to be zero) $\star$	**0.342**	**0.816**
CS1 (mean spreads) ˆ	0.027	0.31
CS2 (mean parameters) ˆ	-0.037	0.008
CS3‡ (negatives to be zero)	**0.322**	0.539
AR‡ (negatives to be zero)	0.269	0.697
ROLL	0.063	0.364
Combination1 (BHL1+SHL2)/2	-0.02	0.061
Combination2 (SHL2+CS2)/2	-0.019	0.292
Combination3 (CS1+BHL1)/2	-0.018	-0.042
Combination4 (BHL1+BHL2)/2	-0.046	-0.109

This table reports the average time-series and cross-sectional correlation of the estimators and the combinations.

Tick by tick quoted data are used to generate the monthly quoted spread.

Highest two correlation coefficients are made bold.

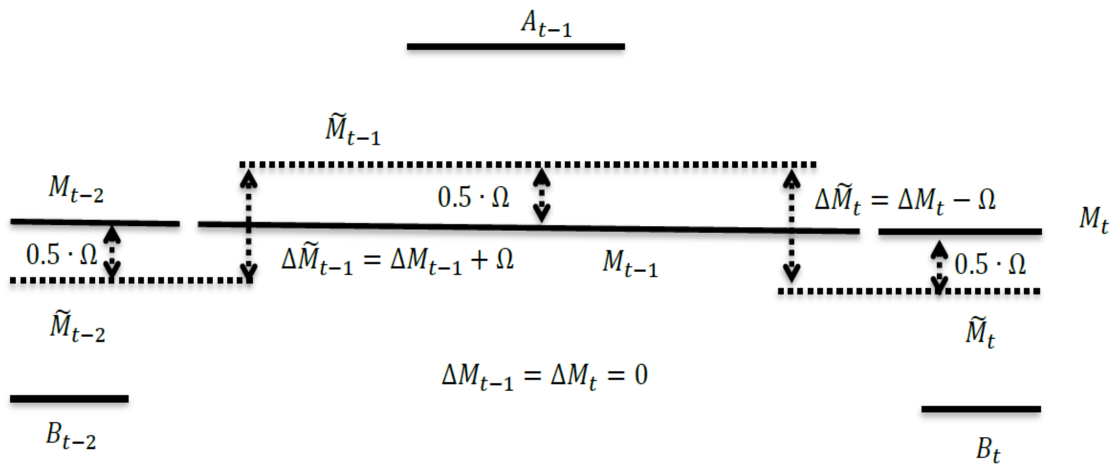

Figure 1: The Conjecture of the Spread

Source Bleaney and Li (2016)

Figure 1 outlines the reasoning underpinning this proposition where for the purposes of economy we hold that the mid-price is fixed. The conjectural spread ($\widetilde{SP}_i$) is less than the true spread. This allows us to estimate the conjectural mid-price $\widetilde{M}$; this is represented by the dotted line in Figure 1, and the true mid price and transaction price are both represented by unbroken lines. Also in Figure 1, A and B denote observed ask and bid prices, whereas M is the unobserved true mid-price. In addition, Δ is taken to be the first-order difference operator and Ω denotes the conjectural error.

At any one point we can only observe one price, either the bid or ask. In Figure 1, three periods are displayed. In the period labelled $t-2$, the bid price is recorded and in period labelled $t-1$, the ask price is observed. In period $t-2$, the conjectural spread is lower than the true spread and the conjectural mid-price error is -0.5Ω , which is less than the true value. In period $t-1$, the conjectural mid-price error is 0.5Ω, therefore this is greater than the true one. In the intervening period between $t-2$ and $t-1$, the direction of the trade shifts from sell to buy, and because of the conjectural error, we overestimate the mid-price return, formally we express this as:

$$\Delta\widetilde{M}_{t-1} = \Delta M_{t-1} + \Omega = \Omega$$

In Figure 1, the hypothetical example shows that the variance of mid-price returns equates to zero because returns remain fixed. However the variance of conjectured mid-price returns is greater than zero. The reason for this is that in the case where the spread is underestimated, the conjectured mid-price fluctuates more than its true counterparts.

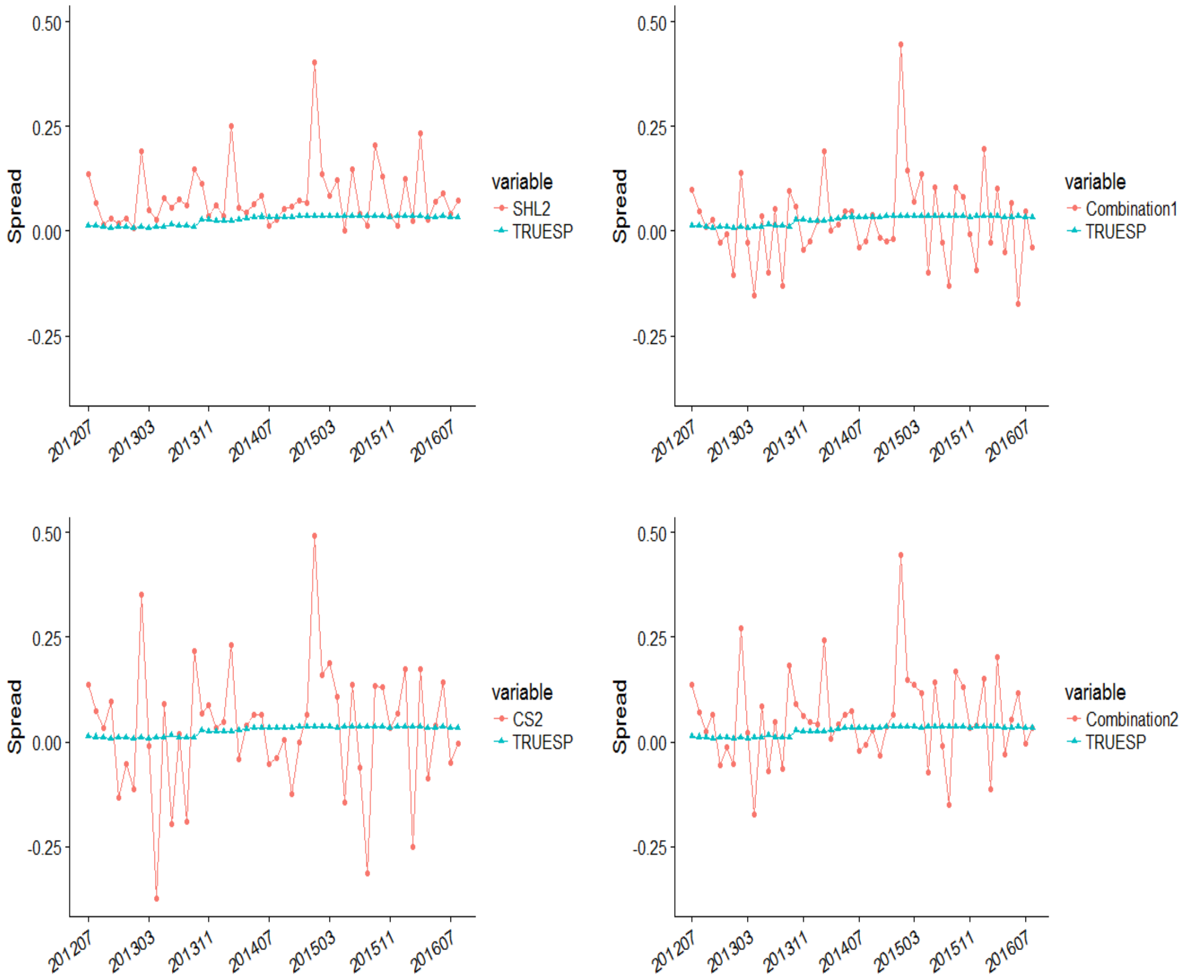

Figure 2: Monthly true and estimated spread, USD/JPY

The graphs above display estimates together with true values for a spread over a 50 month period from July 2012 to August 2016. The currency pair chosen for the illustration without losing generality is USD/JPY. True spreads (TRUESP) data are monthly average closing spreads taken from DataStream.

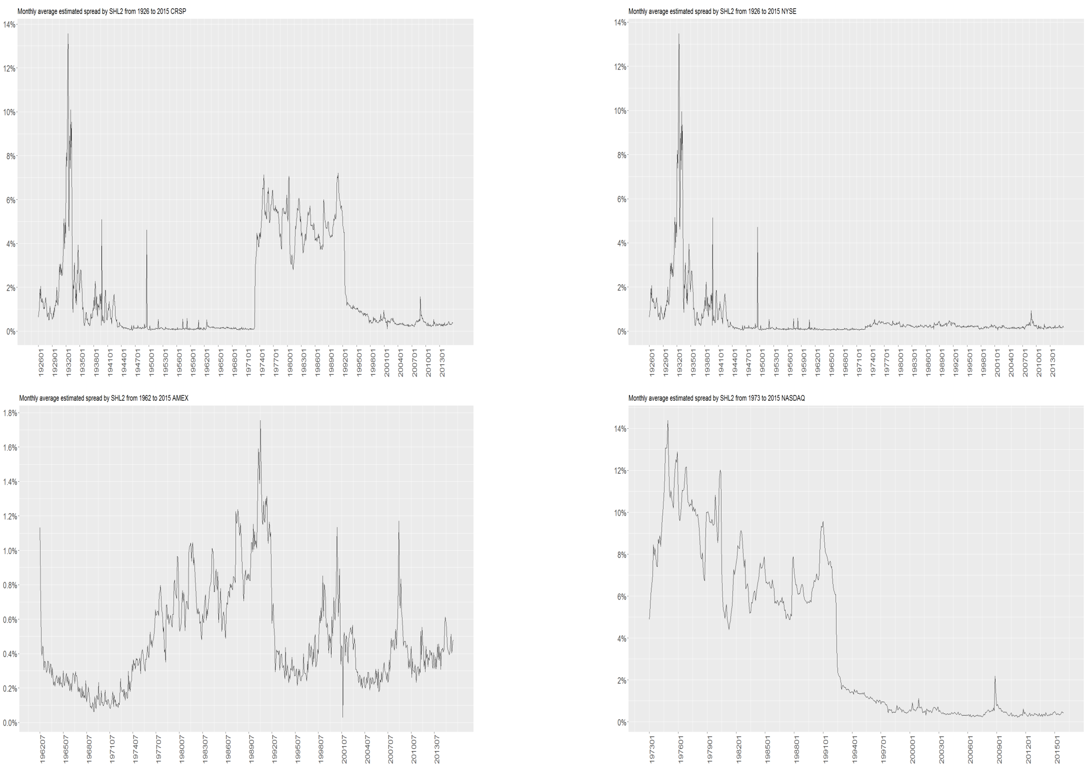

Figure 3: Monthly average estimated spread by SHL2 from 1926 to 2015, CRSP

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the New York Stock Exchange American Stock Exchange and Nasdaq on a monthly basis from January 1926 to December 2015. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from CRSP.

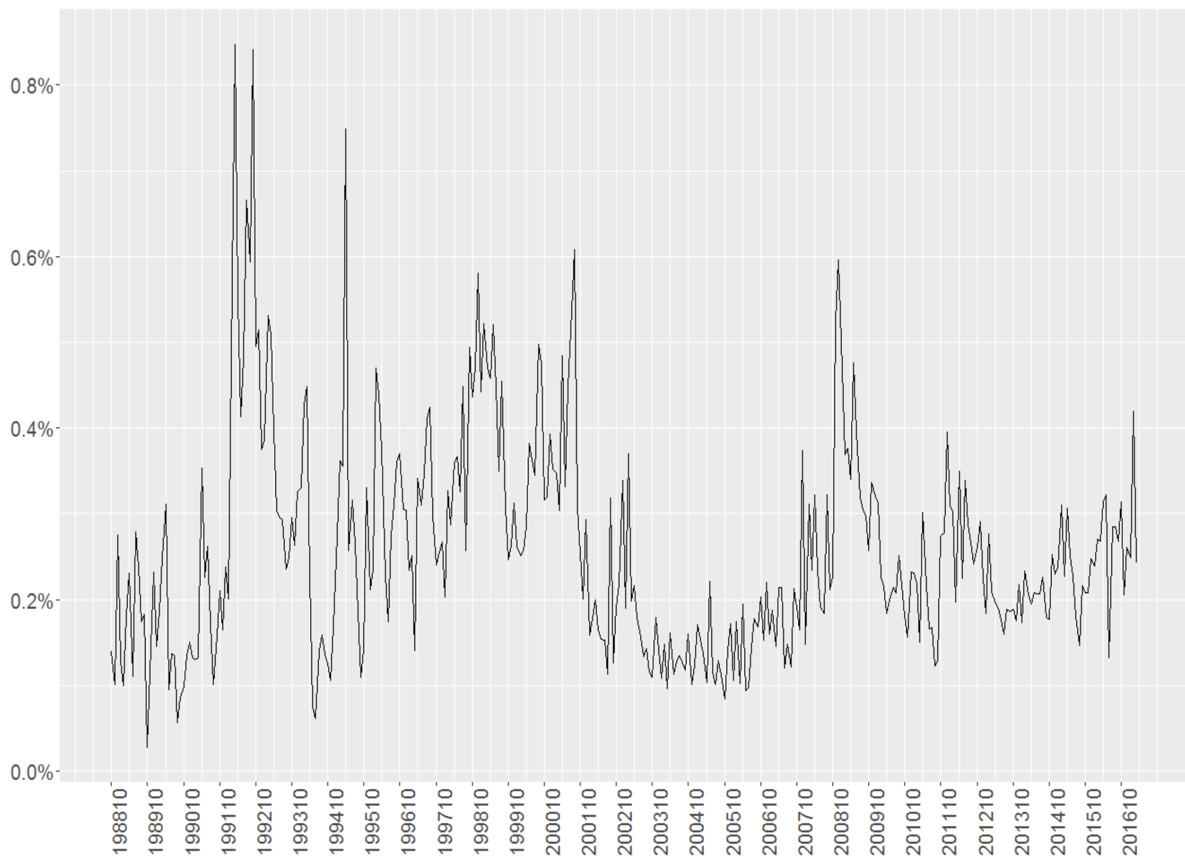

Figure 4: Monthly average estimated spread by SHL2 from 1990 to 2017, UK

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the London Stock Exchange on a monthly basis from October 1988 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with at least 16 daily spread observations within the month. All data is taken from Bloomberg.

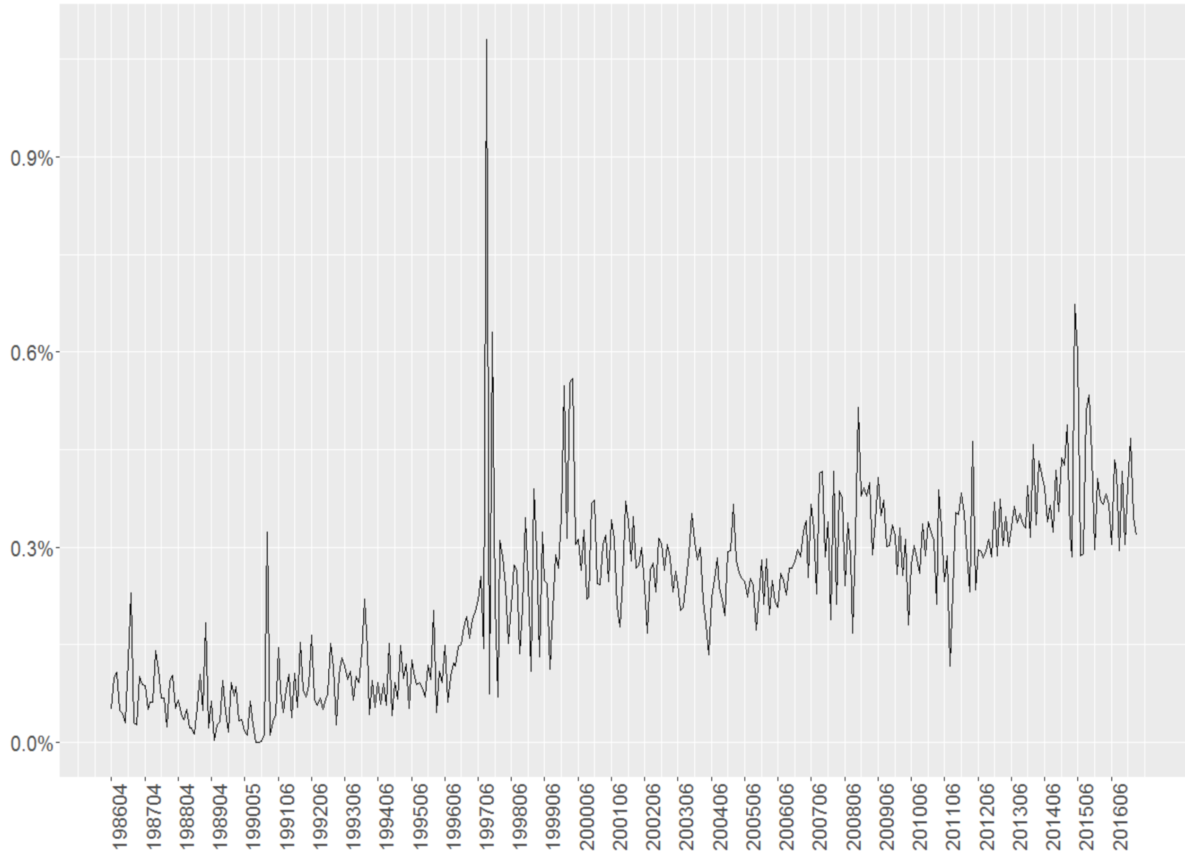

Figure 5: Monthly average estimated spread by SHL2 from 1986 to 2017, Hong Kong

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the Hong Kong Stock Exchange on a monthly basis from April 1986 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from Bloomberg.

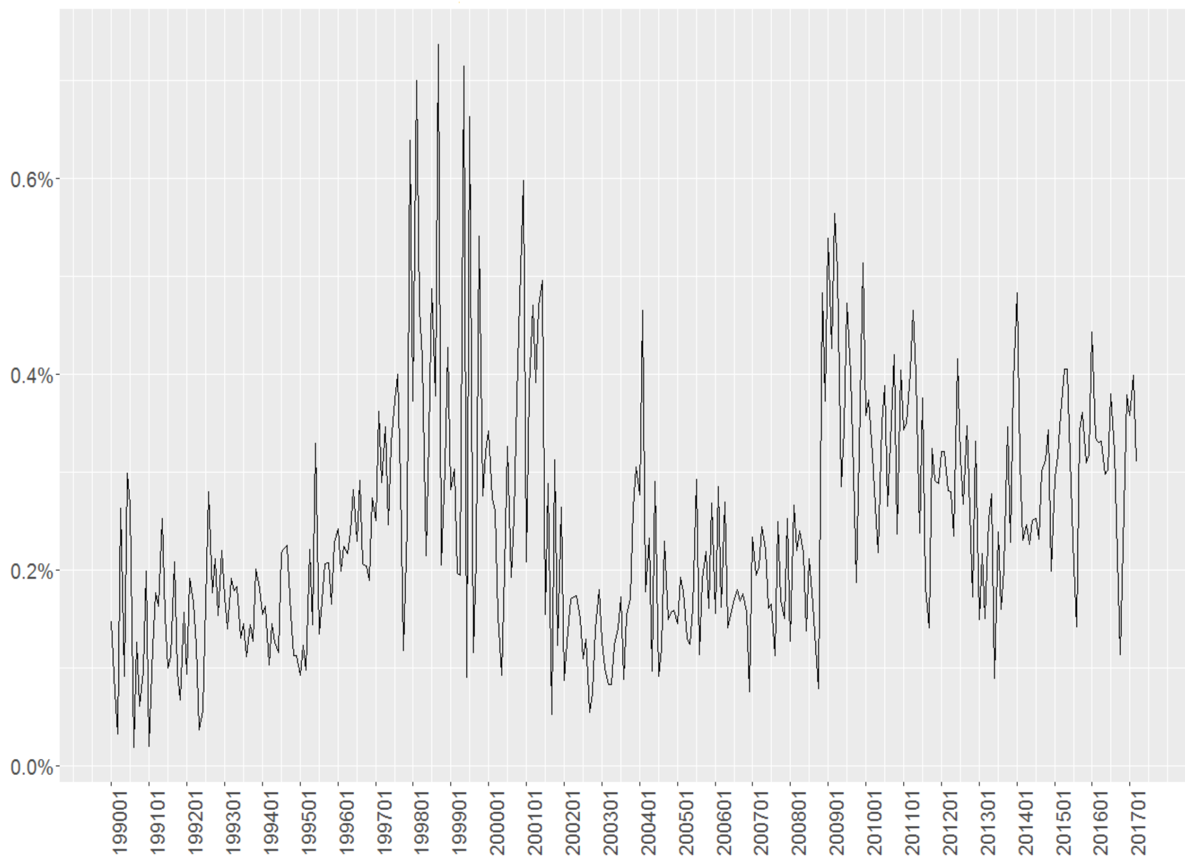

Figure 6: Monthly average estimated spread by SHL2 from 1990 to 2017, Thailand

Depicted here is SHL2 estimated bid-ask spreads for all stocks listed on the Stock Exchange of Thailand on a monthly basis from January 1990 to March 2017. The figure plots the monthly equally weighted average spread of all stocks with each recording at least 16 daily spread observations within the month. All data is taken from Bloomberg.