The performance of estimators of the bid-ask spread under less than ideal conditions

Michael Bleaney and Zhiyong Li *

University of Nottingham

forthcoming Studies in Economics and Finance

Abstract

Purpose: The bid-ask spread is important for many reasons. Because spread data are not always available, many methods have been suggested for estimating the spread. Existing papers focus on the performance of the estimators either under ideal conditions or in real data. The gap between ideal conditions and the properties of real data is usually ignored. The consistency of the estimates across various sampling frequencies is also ignored. This paper investigates the performance of estimators of the bid-ask spread in a wide range of circumstances and sampling frequencies.

Design: The estimators and the possible errors are analysed theoretically. Then we perform simulation experiments, reporting the bias, standard deviation and root mean square estimation error of each estimator. More specifically, we assess the effects of the following factors on the performance of the estimators: the magnitude of the spread relative to returns volatility, randomly varying of spreads, the auto-correlation of mid-price returns, and mid-price changes caused by trade directions and feedback trading.

Findings: The best estimates come from using the highest frequency of data available. The relative performance of estimators can vary quite markedly with

*Corresponding Author: Professor Michael Bleaney, School of Economics, University of Nottingham, University Park, Nottingham NG7 2RD, England; Tel +44 115 95 15464; Email: michael.bleaney@nottingham.ac.uk

the sampling frequency. In small samples, the standard deviation can be more important to the estimation error than bias; in large samples, the opposite tends to be true.

Originality: There is a conspicuous lack of simulation evidence on the comparative performance of different estimators of the spread under the less than ideal conditions that are typical of real-world data. This paper aims to fill this gap.

Keywords: Bid-ask Spread, Feedback Trading, Estimation
JEL: G10

1 Introduction

The spread between ask and bid prices is of interest for a number of reasons: first, it is a useful measure of market participants' trading costs and thus a widely used proxy for market liquidity (e.g. Mancini et al. 2013; Banti et al. 2012); second, it is one of the microstructure noises that cause the observed price series to deviate from the random walk properties assumed by the efficient market hypothesis; and third, it can influence measures of market volatility (e.g. Bandi and Russell 2006).

Because spread data are not always available, many methods have been suggested for estimating it (even if price quotes are available simultaneously for purchases and sales, so that the quoted spread is known, actual transaction prices may differ from quotes, so that the true spread still needs to be estimated). The performance of estimators is generally assessed either by simulation experiments (e.g. Corwin and Schultz 2012) or on real data where the spread is known (ap Gwilym and Thomas 2002; Anand and Karagozoglu 2006; Goyenko et al. 2009; Holden 2009). Real data have features that are likely to affect the estimate of the spread. These features tend to be ignored in simulation experiments, which concentrate on "ideal" conditions where the spread is fixed and mid-price returns and order flows are random. In reality spreads vary with trading volume and size (e.g. McInish and Wood 1992, Chan et al. 1995). Mid-price returns normally exhibit negative autocorrelation (e.g. Goodhart et al. 1996; Daníelsson and Payne 2002), and may be influenced by order flows because of inventory control (e.g. Stoll 1978; Amihud and Mendelson 1980; Ho and Stoll 1981; Bessembinder 1994; Lyons 1995) and asymmetric information costs (e.g. Glosten and Milgrom 1985; Kyle 1985). Order flows may be positively autocorrelated because of hot potato trading or herding (e.g. Sias and Starks 1997; Sias 2004; Lyons 1997; Evans and Lyons 2002; Berger et al. 2008), and may also be affected by mid-price returns, a phenomenon known as feedback trading (e.g. De Long et al. 1990, Hasbrouck 1991, Nofsinger and Sias 1999).

The purpose of the present paper is to fill this gap in the literature by performing simulation experiments that replicate these features of the real data one by one, in order to understand their effect on the absolute and relative performance of different estimators of the spread.

The effect of the sampling frequency on the comparative performance of estimators

is investigated. Lower-frequency data can be obtained from higher-frequency data, and some estimators are explicitly designed for low-frequency data.

A further important aspect is what might be called the signal-to-noise ratio. The spread (the signal) is harder to measure accurately when it is smaller relative to the volatility of the mid-price (the noise). Spreads are much smaller for frequently traded assets than for less frequently traded ones. Since by definition the former dominate trading activity in the market, measuring their spreads accurately is of particular interest. Moreover the signal-to-noise problem is related to the frequency with which the data are sampled. At low frequencies, the signal-to-noise ratio is smaller than at high frequencies, which may affect the performance of low-frequency estimators such as that of Corwin and Schultz (2012). Some estimators might be more sensitive to this problem than others (as we show).

We have chosen four widely used estimators of the bid-ask spread: Roll (1984), Huang and Stoll (1997), Hasbrouck (2004, 2009) and Corwin and Schultz (2012) (referred to as *Roll*, *HS*, *Hasbrouck* and *CS* respectively). The CS estimator is simple to calculate and its data requirements are low. Only the daily high-low price ratio, which is available even in some historical data, is needed. Because the CS estimator is very new, it has been relatively little studied. The Roll estimator is the most widely used, since it requires only price and not transactions data, and has been extended in later research (e.g. Choi et al. 1988, Stoll 1989, George et al. 1991 and Hasbrouck 2004, 2009). Where relevant we also discuss these extensions of the Roll estimator. The Hasbrouck estimator is the latest development of the Roll family of estimators. The HS estimator requires data on trade directions as well as prices (i.e. whether the observed price is a buy or a sell). Obviously, this information greatly improves the accuracy of spread estimation, but it is often unavailable. Unlike the Roll estimator, the HS estimator allows for the possibility that order flows affect subsequent prices.

Our research is related to recent work by Lin (2013), who shows that the accuracy of the Corwin-Schultz estimator increases with both the size of the spread and transaction frequency, and decreases with price volatility. Lin (2013) does not analyse the effects of sampling frequency, time-varying spreads, feedback trading etc., nor does he compare the performance of the Corwin-Schultz estimator with other estimators, as we do here.

Our results show that the chief merit of the CS estimator is that it has a relatively low

standard deviation in low-frequency data. Although it tends to be biased, in many circumstances this consistency makes it the most reliable estimator when only low-frequency data are available. On the other hand, the picture is very different when high-frequency price data are available. Then the Roll estimator is superior to the CS estimator, and the HS estimator generally outperforms the Roll estimator if trade direction data are available.

The rest of the paper is organised as follows. Section Two introduces the estimators and discusses whether they are likely to be biased under various departures from ideal conditions. Section Three reports simulation evidence for sampling frequencies ranging from one minute to 24 hours. The simulations show the average estimation error as well as the bias of each estimator for each experiment. Section Four concludes.

2 Estimators and Errors

In this section we discuss theoretically how various departures from ideal conditions are likely to affect different estimators of the spread. The departures from ideal conditions considered are: time-varying spreads, autocorrelated mid-price returns, autocorrelated trade directions, the price impact of order flows, and feedback trading.

The following equation describes the effective spread,

$$effective\ spread = 2 \cdot |transaction\ price - mid\ price|$$

where mid-price $= 0.5 \cdot (ask + bid)$. Let p_t be the transaction price at time t. It equals the ask/bid price if a buy/sell order is executed,

$$p_t = \begin{cases} ask_t & Buy\ order \\ bid_t & Sell\ order \end{cases} \tag{1}$$

Observed prices can be divided into two parts. One is the bid-ask spread and the other is the unobserved mid-price. Formally, the price is given by,

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

where p_t is the observed price and M_t is the mid-price. SP is the effective bid-ask spread, and BS is the trade indicator which shows the direction of the trade.

$$BS_t = \begin{cases} 1 & buy\ order \\ -1 & sell\ order \end{cases} \tag{3}$$

By taking the first-order difference of the equation above we get an expression for the price return.

$$\Delta p_t = \Delta M_t + \frac{SP}{2}(BS_t - BS_{t-1}) \tag{4}$$

where Δ is the first-order difference operator. It suggests that the spread will enlarge (reduce) the observed return when the change in trade direction has the same (opposite) sign as the mid-price change. If the trade direction does not change ($BS_t - BS_{t-1} = 0$), the observed return is equal to the mid-price change.

The Roll Estimator

Roll (1984) obtains the covariance of price returns from the equation above. By assuming that the mid-price and the trade direction each follow a random walk and spreads are fixed, the Roll estimator is given by:

$$SP = 2\sqrt{-Cov(\Delta p_t, \Delta p_{t-1})} \tag{5}$$

When its assumptions are not valid, the Roll estimator is no longer unbiased. Therefore, it is of interest to evaluate the influence of these factors on the Roll estimator. The errors of the Roll estimator have been well studied, so we only give a very brief introduction.

Error one: the Roll estimator assumes that the spread is fixed. The Roll estimator would overestimate the mean spread when the spread is time-varying.

Error two: the Roll estimator assumes that ΔM_t is independently and identically distributed (iid). The error is positive if the mid-price returns are negatively autocorrelated and vice versa. George et al. (1991) introduce a modified Roll model which is unbiased when the mid-price is autocorrelated.

Error three: when trade directions are autocorrelated, the Roll model is biased. Choi et al. (1988) modify the Roll model to overcome the issue.

Error four: when the inventory control (IC) and asymmetric information (AS) components of the spread are significant, mid-price returns are influenced by past order flows. In these circumstances, the Roll model underestimates the true spread, for it only considers the order-processing element. Stoll (1989) analyses the components of the spread using Roll's framework and obtains an equivalent result. The model cannot be estimated using transaction data only.

Error five: Harris (1990) gives the expression for the Roll estimator when the sample is finite. The bias is a decreasing function of the sample size, and when the sample size is infinite, the Roll estimator is unbiased. Harris (1990) suggests that when the volatility of transaction prices is high, the Roll estimator will underestimate the spread, and requires more observations to overcome the error.

Error six: when there is feedback trading, the Roll estimator is biased. The existence of feedback trading suggests that order flows (trade directions) are influenced by past mid-price returns. The simplest feedback trading is that a trader decides whether to buy or sell according to the most recent mid-price return. In that case, the error is given by:

$$\begin{aligned} Error = 2 \cdot \sqrt{-Cov(\Delta p_t, \Delta p_{t-1})} + Cov(\Delta M_{t-1}, BS_{t-1}) \\ -\sqrt{\left[Cov(\Delta M_{t-1}, BS_{t-1})\right]^2 - 4Cov(\Delta p_t, \Delta p_{t-1})} \end{aligned} \tag{6}$$

The error is positive when there is positive feedback trading and vice versa.

The Huang-Stoll Estimator

Huang and Stoll (1997) relax the assumption that the mid-price follows a random walk; instead it is assumed that the mid-price is affected by two factors. One is the fundamental value and the other one is the inventory level. It is also assumed that the two factors have equal weights. The equation in Huang and Stoll's model is given by:

$$\Delta p_t = \tfrac{SP}{2} BS_t + (\alpha + \beta - 1)\tfrac{SP}{2} BS_{t-1} - \alpha \tfrac{SP}{2}(1 - 2\theta) BS_{t-2} + \epsilon_t \tag{7}$$

where α and β represent the weights of the asymmetric information and inventory control components of the spread respectively. θ is the probability of order reversal. ϵ_t is a random shock. θ can be estimated from,

$$BS_{t-1} = (1 - 2\theta) BS_{t-2} + \varepsilon_t \tag{8}$$

The generalized method of moments is applied to estimate the two equations simultaneously.

When trade directions are correlated with mid-price returns (feedback trading), endogeneity will cause the HS model to be biased. The error should be:

$$Error = 2 \cdot \mathrm{cov}(BS_t, \epsilon_t) \tag{9}$$

Equation (9) suggests that when there is positive feedback trading, the HS estimator will overestimate the true spread and vice versa.

The Hasbrouck Estimator

Hasbrouck (2004, 2009) estimates Equation (4) using the Gibbs sampler. The Hasbrouck model requires transaction prices only. Similar to the Roll estimator, the Hasbrouck model will be influenced by ICAS components, auto-correlated mid-price and feedback trading.

The Corwin-Schultz Estimator

Corwin and Schultz's (2012) spread estimator uses the daily high-low prices to estimate the spread. The following assumptions are made: the mid-price follows a random walk; the spread is fixed; and the highest (lowest) mid-price corresponds to the highest (lowest) transaction price and the buy (sell) order.

From equation (2), the observed high (low) price is assumed to be the highest (lowest) mid-price plus (minus) half of the spread.

$$H_t^O = TH_t^M + \frac{SP}{2} \tag{10}$$

$$L_t^O = TL_t^M - \frac{SP}{2} \tag{11}$$

where H_t^O is the logarithmic observed daily high price and L_t^O is the logarithmic observed daily low price, TH_t^M is logarithmic daily high mid-price and TL_t^M is the logarithmic daily low mid-price. From equations above, one can obtain:

$$H_t^O - L_t^O = TH_t^M - TL_t^M + SP \tag{12}$$

The equation suggests that the high-low ratio of transaction prices is the summation of the high-low ratio of mid-prices and the spread. Squaring both sides of the equation, we have:

$$\left(H_t^O - L_t^O\right)^2 = \left(TH_t^M - TL_t^M\right)^2 + 2\left(TH_t^M - TL_t^M\right) \cdot SP + SP^2 \tag{13}$$

Equations (13) and (12) describe the basic relationship between the high-low ratios of transaction prices and mid-prices and the spread. According to equation (13), the high-low ratio over two-period can be written as:

$$\left(H_{t,t+1}^O - L_{t,t+1}^O\right)^2 = \left(TH_{t,t+1}^M - TL_{t,t+1}^M\right)^2 + 2\left(TH_{t,t+1}^M - TL_{t,t+1}^M\right) \cdot SP + SP^2 \tag{14}$$

where subscript $(t, t+1)$ represents the value is over a two-day interval.

The CS estimator assumes that the movement of mid-prices is a Wiener process, and that the daily high price is an ask price and the daily low price is a bid price. Parkinson (1980) and Garman and Klass (1980) discuss using high-low ratios as a measure of volatility under these assumptions. The CS estimator is based on the insight that the relationship between high-low ratios measured over different lengths of time is influenced by the spread. They use spreads over one-day and two-day intervals for this purpose. A spread estimate for each two-day interval is obtained by using the high-low ratio for each of the days individually and over the whole two days. The system of the CS estimator can be written as:

$$2 \cdot k_1\sigma^2 + 4 \cdot k_2 \cdot \sigma SP + 2SP^2 - \beta = 0 \tag{15}$$

$$2 \cdot k_1\sigma^2 + 2\sqrt{2} \cdot k_2 \cdot \sigma SP + SP^2 - \gamma = 0 \tag{16}$$

where k_1 and k_2 are the constants suggested by Parkinson (1980) and Garman and Klass (1980), and following Corwin and Schultz (2012), we assume that $k_1 = {k_2}^2$; σ and σ^2 are the standard deviation and variance of the mid-price returns; β is the square of the summation of two adjacent daily high-low ratios; and γ is the square of the high-low ratio in the same two-day interval.

$$\beta = E\left\{\sum_{J=0}^{1}\left(H_{t+J}^{O} - L_{t+J}^{O}\right)^2\right\}; \quad \gamma = \left(H_{t,t+1}^{O} - L_{t,t+1}^{O}\right)^2 \tag{17}$$

The solutions to the equation system are then given by:

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

where

$$\alpha = \frac{\sqrt{2\beta} - \sqrt{\beta}}{3 - 2\sqrt{2}} - \sqrt{\frac{\gamma}{3 - 2\sqrt{2}}} \tag{18}$$

When the spread is small, $SP \approx \alpha$. We may therefore use equation (18) to estimate the spread.

This procedure yields a spread estimate for each two-day interval; the mean spread estimate is obtained from the average of the solutions for the n (overlapping) two-day intervals in the sample:

$$\overline{SP} = \frac{1}{n}\sum_{t=1}^{n} SP \tag{19}$$

The CS estimator is biased for two reasons. Firstly, some assumptions may not be valid. These errors can be called the spread errors. Secondly, the average of spreads is

not an unbiased estimator of the expectation of spreads due to the non-linear equation system (Corwin and Schultz 2012). In other words, because of non-linearity, even zero-mean errors cannot be completely eliminated by increasing the sample size. This second type of error can be called equation errors. Formally, the solutions are given by:

$$\overline{SP} = \lim_{n\to+\infty} \frac{1}{n} \sum_{t=1}^{n} \left[\frac{\sqrt{2(\beta_t + e_{\beta,t})} - \sqrt{\beta_t + e_{\beta,t}}}{3-2\sqrt{2}} - \sqrt{\frac{\gamma_t + e_{\gamma,t}}{3-2\sqrt{2}}} - e_{sp} \right] \tag{20}$$

where e_{sp} is a spread error and e_γ and e_β are equation errors. It is straightforward that the solutions will reflect the true spread when $E(e_{sp}) = 0$, but will not converge to the true spread even when $E(e_{\beta,t}) = E(e_{\gamma,t}) = 0$. Equation (20) is the general form of the estimation of σ or the spread. The following paragraphs will discuss specific errors one by one.

Error one (a spread error): the CS estimator assumes that the mid-price of the highest (lowest) transaction price corresponds to the highest (lowest) mid-price. The assumption is not always valid when the standard deviation of mid-price is much larger than the spread. In these circumstances, the true highest (lowest) mid-prices may be not used but the less (greater) ones, and as a result of which the true spread is underestimated. A formal discussion appears in the Appendix.

Error two (a spread error): the estimator also assumes that the highest (lowest) transaction price normally happens on a buy (sell) order. The assumption is not true when the number of trades in the interval is small. Formally, when the assumption is not true, the transaction prices are given by:

$$\begin{aligned} H_t^O &= TH_t^M - \tfrac{SP}{2} \quad (a) \\ L_t^O &= TL_t^M + \tfrac{SP}{2} \quad (b) \end{aligned} \tag{21}$$

The equations above suggest that the spread will be underestimated if either (a) or (b) happen. The error is either equal to or twice as large as the spread. The probability of events (a) and (b) happening is negatively correlated with the number of observations in an interval. Thus the error is also negatively correlated with the number of observations in an interval. One can easily show that if the probability of events (a) or (b) happening is $1-\eta$ where $0 \leq \eta \leq 1$, the error is given by:

$$e_o = 2\eta \cdot SP \tag{22}$$

Error three (an equation error): equation (16) is not always valid. Equation (16) builds a link between the one-period volatility and the two-period volatility basing on Parkinson's (1980) volatility estimator. See the Appendix for further discussion.

Error four (an equation error): the approximation that $\sqrt{k_1} = k_2$ can also introduce an error to the system. Formally, the error is given by,

$$e_a = 2k_1\sigma^2 - 2k_2^2\sigma^2 = 0.452\sigma^2 \tag{23}$$

e_a influences both e_β and e_γ. It is straightforward that the error is positively correlated with the volatility of mid-price returns.

Error five: the CS estimator is derived under the assumption that the mid-price follows a Brownian motion. When the mid-price is autocorrelated, Parkinson's (1980) estimator, on which the CS estimator is based, is no longer unbiased, and as a result, the CS estimator could underestimate the variance of the price returns if the mid-price is negatively correlated and thus overestimate the spread.

Error six: as with the Roll estimator, the time-varying spread can also influence the accuracy of the estimator because of the associated non-linearity.

Error seven: as with the Roll estimator, the CS estimator does not consider the existence of the price impact components of the spread. Therefore, the CS estimator will underestimate the true spread when these components are not zero.

Furthermore, because of the non-linearity, the joint effect of errors could be several times as big as the summation of single errors. Most of the errors of the CS estimator increase with mid-price returns volatility, as we demonstrate in the simulations.

3 Simulation Experiments

The aim of this section is to empirically compare the estimators and analyse the errors which are discussed theoretically in the previous section, using simulation experiments. Sampling frequencies ranging from one minute to 24 hours are considered. More specifically, we assess the effects of the following factors on the performance of the estimators: the magnitude of the spread relative to the returns volatility, the variation of spreads, the autocorrelation of mid-price returns, and mid-price changes caused by trade directions and feedback trading. For each experiment we report the bias of each estimator (the

deviation of the mean from the true value), its standard deviation about the mean, and the root mean square estimation error (the standard deviation about the true value).

Most of our simulations are based on a spread size and mid-price returns volatility that are typical of frequently traded currencies in the foreign exchange market, as reported by Lyons (1995). One advantage of focusing on the foreign exchange market is that the market is continuously open, so we do not have to consider issues of news flow outside trading hours. In some simulations we also consider the case of much bigger spreads, to illustrate the effect of spread size. We ignore idiosyncratic features of particular markets, such as the minimum tick rule of the New York Stock Exchange.

3.1 Comparison Procedure

The general principle of the comparison is that, using simulation experiments, we start from the ideal conditions for the estimators and change one condition each time, so that we can identify the influence of the condition. The choice of conditions is according to the theoretical analysis in the previous section. The cases are listed as follows:

1) Ideal case with a big spread

2) Ideal case with a smaller bid-ask spread close to that in the foreign exchange market

3) Time-varying spread

4) Price impact: the inventory control and asymmetric information costs

5) Negatively auto-correlated mid-price returns

6) Positively auto-correlated trade directions

7) Feedback trading

We first consider the ideal case with random mid-price returns and trade directions, and with spreads that are large and fixed over time, so that all the assumptions of the estimators are satisfied. The mid-price returns volatility used is close to the real volatility in the foreign exchange market, but the spread is much larger. These are the conditions most favourable for spread estimation. Next, we study the case of smaller bid-ask spreads, keeping everything else the same. The smaller bid-ask spread used in this case is close to the real spread in the foreign exchange market.

Then we consider cases where the assumptions underlying the estimators are not met. First, we allow the bid-ask spread to be time-varying. Next, we allow mid-price returns

to be influenced by trade direction, which suggests that there are inventory control and asymmetric information components of the spread. Bessembinder (1994) and Lyons (1995) provide evidence for the presence of these costs in the FX market.

In the fifth simulation, we allow mid-price returns to be negatively auto-correlated, as tends to happen in the foreign exchange market (Goodhart et al. 1996; Daníelsson and Payne 2002). In the sixth simulation, auto-correlated trade directions are considered. The autocorrelation of trade direction could be caused by herding behaviour, hot-potato trading, clustering or other reasons. Finally, the case of positive feedback trading is considered. Feedback trading suggests that trade directions might be influenced by price returns. De Long et al. (1990) introduce a theoretical model of positive-feedback trading. Hasbrouck (1991), Nofsinger and Sias (1999) and Daníelsson and Love (2006) show the existence of feedback trading in the stock and foreign exchange markets.

3.2 Settings of simulation experiments

This section introduces the general settings of simulation experiments. There are 1000 replications simulated for each case. There are 432000 periods in a replication. Let one period represent one minute, with exactly one trade per minute. The market is continuously open. Thus there are 300 trading days (1440 minutes and 1440 trades per day). For each replication, data are considered in various sampling periods: tick-by-tick, and five-minute, fifteen-minute, 30-minute, one-hour, four-hour, 12-hour and 24-hour. Thus, there are eight subgroups for each replication. To obtain five-minute and longer sampling-period data, we use only the closest observation to the sampling time in the generated tick-by-tick data. Only trade direction and transaction prices are assumed to be observed. Data are generated according to the following system. Trade direction has two possible values 1 and -1. Trade direction is either random (if $\varphi = \psi = 0$), autocorrelated (if $\psi = \phi = 0$) or a function of mid-price returns (if $\varphi = \phi = 0$). Formally, the trade direction series is given as follows.

$$\begin{aligned} &BS_t = \varphi F(BS_{t-1}) + \psi[I_t(\Delta M_t) - 0.5] \cdot 2 + \phi(\varpi_t - 0.5) \cdot 2 \\ &\varphi = 0 \; or \; 1; \psi = 0 \; or \; 1; \phi = 0 \; or \; 1 \\ &\varphi + \psi + \phi = 1 \end{aligned} \tag{24}$$

where BS_t is the trade direction, $F(BS_{t-1})$ is a function of the past trade direction, which suggests that the trade direction is autocorrelated, and the outcome of $I_t(\cdot)$ is a binomial random variable which is 1 or 0. Then $[I_t(\cdot) - 0.5] \cdot 2$ is 1 or -1. $I_t(\cdot)$ is influenced by the past mid-price return (assume the trader observes the mid-price return first and then places the order, thus ΔM_t is the past return for the trader at period t), which suggests the existence of feedback trading. The function $I_t(\cdot)$ reflects the following relationship between trade directions and past mid-price returns.

$$I_t(\Delta M_t) \sim \begin{cases} B(1,\kappa) & if \quad \Delta M_t > 0 \\ B(1,1-\kappa) & if \quad \Delta M_t < 0 \end{cases} \tag{25}$$

where $B(1,\kappa)$ is a binomial distribution with one trial and κ probability. When $\kappa = 0.5$, there is no feedback trading, and when $\kappa > 0.5$, there is positive feedback trading and vice versa. ϖ_t is a binomial random variable, which follows a binomial distribution with one trial and 50% probability i.e. $B(1, 0.5)$. Then $(\varpi_t - 0.5) \cdot 2$ is 1 or -1. It suggests that trade directions are drawn from a binomial distribution randomly and both the buy and sell orders carry the same weight in the series. φ, ψ and ϕ are weighting coefficients all of which have two possible values 0 or 1. Let the sum of these coefficients equal one, which suggests that only one of the coefficients could be one. The setting makes sure only one factor is considered each time, so we can identify the influence of the factors separately.

Mid-price returns are generated using the following equation,

$$\begin{aligned} &\Delta M_t = \tau\, \xi \Delta M_{t-1} + \omega\chi\, BS_{t-1} \cdot \tfrac{SP_t}{2} + \varepsilon_t \\ &\tau = 0 \; or \; 1; \omega = 0 \; or \; 1; \\ &\tau + \omega \leq 1 \end{aligned} \tag{26}$$

where ξ describes the autocorrelation of mid-price returns. ε_t follows a normal distribution with zero mean and standard deviation σ; SP_t is the bid-ask spread that follows a normal distribution $N(\mu, \varsigma^2)$, where μ is the mean and ς is the standard deviation. SP_t could be negative, which does not influence the estimators mathematically. When $\varsigma = 0$, spreads are fixed. χ is the fraction of inventory control and asymmetric information components of the spread, and thus $(1 - \chi)$ suggests the order-processing component of the spread. When $\chi = 0$, the order-processing cost is the only component of the spread. When $\xi = 0$ and $\chi = 0$, mid-price follow a random walk process. Let the summation of τ and ω equal to one, which suggests that only one of the coefficients could be one. The setting makes

sure only one factor is considered each time, thus we can identify the influence of the factors separately.

Transaction prices are generated by

$$p_t = M_t + \frac{SP_t}{2} \cdot BS_t \tag{27}$$

where p_t is the transaction price.

3.3 Fixed Spread

In this section, the ideal case for the estimators is considered, where trade directions are random; mid-prices follow a random walk process; the volatility of mid-price returns is small relative to the spread; and the spread is fixed and relatively big. Under these circumstances, both the HS and the Roll estimators are unbiased, and the error in the CS estimator should be small. Formally, let $\varphi = 0$, $\psi = 0$, $\phi = 1$ in equation (24), which suggests that trade directions are random. Let $\tau + \omega = 0$ in equation (26), which suggests that the mid-price follows a random walk process. The standard deviations of mid-price returns is $\sigma = 0.0002$. $\mu = 0.03$, $\varsigma^2 = 0$ which suggests that the spread is a constant. The system is given by,

$$\begin{aligned} BS_t &= 2 \cdot (\varpi_t - 0.5) \\ \varpi_t &\sim B(1, 0.5) \\ \Delta M_t &= \varepsilon_t \\ \varepsilon_t &\sim N(0, 4 \times 10^{-8}) \\ SP_t &= 0.03 \\ p_t &= M_t + \tfrac{SP_t}{2} \cdot BS_t \end{aligned} \tag{28}$$

One thousand replications, each of which has 432000 periods, are generated according to the system above. As mentioned earlier, each replication has eight subgroups according to various sampling periods. Thus, for every subgroup, there are 1000 estimated spreads for each estimator and 1000 standard deviations of mid-price returns.

The results are presented in Table 1. There are four panels which report the summary statistics and the results of the estimators respectively. The columns represent the sampling frequency of the data. The first row shows the average standard deviation of mid-price returns over the sampling interval. The second row shows the ratio of the spread to this, which falls from 150 in tick-by-tick data to 3.96 in 24-hour sampling. Then, for

each of the three estimators, four rows of statistics are shown: (1) the mean spread estimate; (2) the ratio of this estimate to the true value (in this case 0.03); (3) the standard deviation of the spread estimate; and (4) the root mean square error [1] (RMSE), i.e. the standard deviation of the spread estimates about the true value (rather than about the estimated mean). The RMSE captures both the bias in the estimate and its variability.

The Roll estimator is an unbiased estimator of the true spread, according to the first row of panel Roll. The third row shows that the standard deviation of the estimated spreads increases with the time interval between observations, but is at most less than 6% of the spread (in the 24-hour case), which suggests that the results are stable across the replications. Therefore, one can conclude that the Roll estimator works well under the ideal conditions.

The HS estimator is as unbiased as the Roll estimator, but far more accurate: even in the 24-hour case, the standard deviation is only about 0.3% of the spread. As in the case of the Roll estimator, the accuracy declines markedly as the time interval increases.

The Hasbrouck estimator slightly underestimates the spread. In short time intervals, the standard deviation is greater than Roll and HS estimators. When time interval is four hours or longer, the standard deviation is smaller than these estimators.

The CS estimator exhibits a slight downward bias, tending to underestimate the spread by about 1%, but the bias gets smaller at longer time intervals. However a significant feature of the CS estimator is that the standard deviation of the estimates is the smallest of all three; it is about half as great as the HS standard deviation at any given time interval. The CS standard deviation is about one-thirtieth of the Roll standard deviation at high frequencies, increasing to one-fifth at 24 hours.

As the time interval lengthens, the combination of a modest fall in the bias of the CS estimator and (especially) the increases in the standard deviation for all estimators, of which the CS is the lowest, means that the RMSE of the CS estimator increases less rapidly with the time interval than does the RMSE of the Roll and HS estimators. Indeed at 12-hour and 24-hour intervals the CS estimator has a lower RMSE than the Roll or the HS estimator; at one-hour and four-hour intervals the HS estimator has the lowest RMSE,

[1] Some authors allow the spread to vary and calculate a correlation between the true and estimated spread. This correlation will tend to be higher when the standard deviation of spread estimates is lower, but unlike the RMSE, it does not take any account of the bias in the estimates.

followed by the CS estimator; at 15-minute and 30-minute intervals, the CS estimator has the highest RMSE, and the HS estimator the lowest. However, when the CS estimator is relatively good (at low frequencies), its RMSE is still higher than the RMSE of the Roll and HS estimators at high frequencies.

Using the RMSE as the criterion, therefore, this simulation would suggest the following recommendation for ideal conditions:

1) If high-frequency data are available, use the HS estimator if trade direction information exists; otherwise use the Roll estimator.

2) If only low-frequency data are available, use the CS or Hasbrouck estimators even if trade direction information exists.

Table 1: Fixed Big Spread (Spread=0.03)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.200	0.447	0.775	1.10	1.55	3.10	5.36	7.58
Spread/(returns SD)	150	67.1	38.7	27.3	19.4	9.68	5.60	3.96
				Roll 1984				
Estimates $\times 10^{-3}$	30	30	30	30	30	30	30	30
Relative Estimate	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Est-Std $\times 10^{-3}$	0.0520	0.113	0.195	0.283	0.410	0.807	1.47	2.02
RMSE $\times 10^{-3}$	0.0520	0.113	0.195	0.283	0.410	0.807	1.47	2.02
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	30	30	30	30	30	30	30	30
Relative Estimate	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Est-Std$\times 10^{-3}$	0.000621	0.00303	0.00898	0.0183	0.0366	0.136	0.426	0.871
RMSE$\times 10^{-3}$	0.000621	0.00303	0.00898	0.0183	0.0366	0.136	0.426	0.871
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			29.5	29.6	29.6	29.6	29.7	29.8
Relative Estimate			0.983	0.987	0.987	0.987	0.990	0.993
Est-Std$\times 10^{-3}$			0.00558	0.00928	0.0177	0.0722	0.210	0.401
RMSE$\times 10^{-3}$			0.500	0.400	0.400	0.406	0.366	0.448
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	29.8	29.8	29.8	29.7	29.7	29.7	29.9	29.9
Relative Estimate	0.993	0.993	0.993	0.99	0.99	0.99	0.997	0.997
Est-Std$\times 10^{-3}$	0.0919	0.123	0.0629	0.148	0.0708	0.13	0.333	0.663
RMSE$\times 10^{-3}$	0.220	0.235	0.210	0.335	0.308	0.327	0.348	0.670

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is drawn from a binomial distribution, i.e. $BS_t \sim B(1, 0.5)$. The mid-price return is drawn from a normal distribution of which the mean is zero and the variance is 9×10^{-12}, i.e. $\Delta M_t \sim N(0, 9 \times 10^{-12})$. The spread is fixed and equals to 0.0003, i.e. $SP_t = 0.03$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

Mid-price returns SD is the average of the standard deviations of mid-price returns.

Estimates is the average of the estimated spreads.

Relative Estimate represents the average of estimated spreads divided by the true spread. It is one if the estimate equals the true spread.

Est-Std is the standard deviation of the estimated spreads.

RMSE is the Root Mean Square Error.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

3.4 Fixed Small Spread

In this simulation, everything is the same as in the previous section except that the spread is 100 times as small (0.0003 instead of 0.03), and more representative of the foreign exchange market. As before, the standard deviation of one-minute mid-price returns is $\sigma = 0.0002$, so the ratio of the spread to the standard deviation of mid-price returns varies from 1.5 at one minute down to 0.04 at 24 hours. Table 2 shows the results. With such a small spread, it is harder for the estimators to distinguish the spread from the noise. Nevertheless the HS estimator has very little bias ($< 1\%$) up to four hours, but underestimates by nearly 6% on average at 24 hours. Both the Roll and Hasbrouck estimators are unbiased at five minutes and fifteen minutes, but seriously overestimate on average at four hours and longer. The standard deviation of the Hasbrouck estimator is smaller than Roll and the HS estimator at 12 hours or longer. The CS estimator tends to underestimate at higher frequencies (up to four hours) and to overestimate (but by less than the Roll estimator) at twelve and 24 hours. Once again, the CS estimator has the lowest standard deviation at all time intervals.

The general conclusion is the same as for Table 1: the CS estimator has the lowest RMSE at time intervals of four hours or more, but if high-frequency data are available, then the HS or Roll estimator applied at high frequencies has a much lower RMSE than the CS estimator at low frequencies. Table 1 showed that with a relatively large spread the HS estimator had by far the lowest RMSE in high-frequency data. With a much smaller spread (Table 2), the HS estimator still has the lowest RMSE at high frequency, but interestingly the Roll and Hasbrouck estimators are much more competitive.

3.5 The Effect of Sample Size

We do not report a full analysis of the effect of sample size, but we did examine its effects in the case of 24-hour sampling. Table 3 compares the performance of the estimators at 24 hours in various sample sizes (150 days, 300 days and 3000 days). The settings of the simulation experiments are the same as Table 2. The column of 150-days uses the first 150 observations in the last column of Table 2 to estimate the spread. The column of 300-days is the same as the last column of Table 2. The standard deviations of all estimators decrease as the sample size becomes greater. Unlike the HS estimator, the

Roll, and Hasbrouck estimators, the CS estimator does not exhibit consistency as the sample size increases. The drawback of the HS estimator, the big standard deviation in long time intervals, can be overcome by adding more observations. According to the RMSEs, the HS estimator is the best choice even at long time intervals.

Table 2: Fixed Small Spread (Spread=0.0003)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.200	0.447	0.775	1.10	1.55	3.10	5.36	7.58
Spread/(returns SD)	1.5	0.671	0.387	0.273	0.194	0.0968	0.0560	0.0396
				Roll 1984				
Estimates $\times 10^{-3}$	0.300	0.300	0.300	0.292	0.265	0.422	0.921	1.58
Relative Estimate	1.000	1.000	1.000	0.973	0.883	1.407	3.070	5.267
Est-Std$\times 10^{-3}$	0.000886	0.00587	0.0246	0.0729	0.184	0.471	1.05	1.77
RMSE$\times 10^{-3}$	0.000886	0.00587	0.0246	0.0733	0.187	0.487	1.22	2.18
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	0.300	0.300	0.301	0.301	0.298	0.298	0.295	0.283
Relative Estimate	1.000	1.000	1.003	1.003	0.993	0.993	0.983	0.943
Est-Std$\times 10^{-3}$	0.000622	0.00296	0.00898	0.0176	0.0364	0.142	0.421	0.880
RMSE$\times 10^{-3}$	0.000622	0.00296	0.00904	0.0176	0.0365	0.142	0.421	0.880
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			-0.0478	0.00979	0.0740	0.258	0.492	0.732
Relative Estimate			-0.159	0.033	0.247	0.860	1.640	2.440
Est-Std$\times 10^{-3}$			0.00447	0.00806	0.0170	0.0665	0.207	0.397
RMSE$\times 10^{-3}$			0.348	0.290	0.227	0.0787	0.282	0.587
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	0.3	0.3	0.295	0.288	0.275	0.604	1.43	2.31
Relative Estimate	1.000	1.000	0.983	0.960	0.917	2.013	4.767	7.700
Est-Std$\times 10^{-3}$	0.000968	0.00565	0.0257	0.0659	0.109	0.205	0.51	0.77
RMSE$\times 10^{-3}$	0.001	0.006	0.026	0.067	0.112	0.367	1.240	2.152

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is drawn from a binomial distribution, i.e. $BS_t \sim B(1, 0.5)$. The mid-price return is drawn from a normal distribution of which the mean is zero and the variance is 4×10^{-8}, i.e. $\Delta M_t \sim N(0, 4 \times 10^{-8})$. The spread is fixed and equals to 0.0003, i.e. $SP_t = 0.0003$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four-hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

Table 3: Effect of sample size on estimator performance in daily sampling (spread = 0.0003)

	150 days	300 days	3000 days
		Estimates$\times 10^{-3}$	
Roll	1.999	1.58	0.85
Huang & Stoll	0.295	0.283	0.316
Corwin & Schultz	0.750	0.732	0.751
Hasbrouck	2.898	2.31	1.275
		Relative Estimate	
Roll	6.662	5.267	2.843
Huang & Stoll	0.982	0.943	1.053
Corwin & Schultz	2.499	2.440	2.503
Hasbrouck	9.659	7.700	4.250
		Est-std$\times 10^{-3}$	
Roll	2.132	1.77	0.984
Huang & Stoll	1.246	0.880	0.281
Corwin & Schultz	0.567	0.397	0.125
Hasbrouck	1.018	0.77	0.476
		RMSE$\times 10^{-3}$	
Roll	2.726	2.18	1.129
Huang& Stoll	1.246	0.880	0.281
Corwin & Schultz	0.723	0.587	0.468
Hasbrouck	2.790	2.152	1.085

There are 1000 replications. Data of each replication are generated according to the same system as the previous table. Each replication is also sampled into a longer time interval: twenty-four-hour, and only the last observation is kept. Columns 1 reports the estimates using the data of the first 216000 periods in Table 2. Thus, there are 150 days for each replication. Column 2 is the same as the last column in Table 2. There are 300 days for each replication. In Column 3, there are 4320000 periods, each of which represents one minute, in each replication. Thus, there are 3000 days for each replication. The standard deviation of mid-price returns, and the estimated spread are collected.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

3.6 Time-varying Spreads

In this section, instead of a fixed spread of 0.03, as in Table 1, we allow the spread to follow a normal distribution with a mean of 0.03 and a standard deviation of 0.01, i.e. $N\left(0.03, 10^{-4}\right)$. Thus the mean spread is the same as in Table 1, but the actual spread varies randomly over time. Spreads may be higher at certain times either because the market is thin (e.g. in the hours after midnight GMT), or because customers are more anxious to trade large quantities to close their books, and larger trades attract a higher spread (Moulton 2005). The results are shown in Table 4. Any differences between Table 4 and Table 1 are because the spread is now allowed to vary over time. The system is given by:

$$\begin{aligned} BS_t &= 2 \cdot (\varpi_t - 0.5) \\ \varpi_t &\sim B(1, 0.5) \\ \Delta M_t &= \varepsilon_t \\ \varepsilon_t &\sim N(0, 4 \times 10^{-8}) \\ SP_t &\sim N\left(0.03, 10^{-4}\right) \\ p_t &= M_t + \tfrac{SP_t}{2} \cdot BS_t \end{aligned} \tag{29}$$

The results are presented in Table 4. The HS estimator remains unbiased, but its standard deviation is about 50% larger than in Table 1. The Roll estimator is no longer unbiased, and overestimates the mean spread by 20% at all time intervals. As in Table 1, its standard deviation is also larger than that of the HS estimator in all cases. The Hasbrouck estimator overestimates the spread slightly. The CS estimator is the most badly affected by time-varying spreads. It overestimates by 36% at 15-minute intervals, rising to 166% at 24 hours. Although the CS estimator still has the smallest standard deviation, the large bias means that its RMSE is greater than that of the Roll and Hasbrouck at any time interval, whereas in Table 1 it had a lower RMSE than even the HS estimator at twelve and 24 hours. Thus when the spread is known to be significantly time-varying, and only low-frequency price data are available (and no trade direction data), the CS estimator is no longer superior to the Roll and Hasbrouck estimators. When the information about trade direction is not available, the Hasbrouck estimator is the best choice.

Table 4: Time-varying Spreads (mean = 0.03)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.200	0.447	0.775	1.10	1.55	3.10	5.36	7.58
Spread/(returns SD)	150	67.1	38.7	27.3	19.4	9.68	5.60	3.96
				Roll 1984				
Estimates $\times 10^{-3}$	36.1	36.1	36.1	36.0	36.1	36.0	36.2	35.9
Relative Estimate	1.203	1.203	1.203	1.200	1.203	1.200	1.207	1.197
Est-Std$\times 10^{-3}$	0.0553	0.126	0.218	0.300	0.425	0.873	1.56	2.32
RMSE$\times 10^{-3}$	6.10	6.10	6.10	6.01	6.11	6.06	6.39	6.34
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	30	30	30	30	30	30	30	29.8
Relative Estimate	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.993
Est-Std$\times 10^{-3}$	0.0209	0.0488	0.0828	0.124	0.171	0.364	0.748	1.21
RMSE$\times 10^{-3}$	0.0209	0.0488	0.0828	0.124	0.171	0.364	0.748	1.23
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			40.9	50.1	57.7	69.7	76.7	79.8
Relative Estimate			1.363	1.670	1.923	2.323	2.557	2.660
Est-Std$\times 10^{-3}$			0.0385	0.0453	0.0580	0.113	0.254	0.471
RMSE$\times 10^{-3}$			10.9	20.1	27.7	39.7	46.7	49.8
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	31.1	31.1	31.1	31.1	31.1	31.2	31.4	31.5
Relative Estimate	1.037	1.037	1.037	1.037	1.037	1.040	1.047	1.050
Est-Std$\times 10^{-3}$	0.106	0.141	0.143	0.225	0.272	0.582	1.04	1.54
RMSE$\times 10^{-3}$	1.105	1.109	1.109	1.123	1.133	1.334	1.744	2.150

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is drawn from a binomial distribution, i.e. $BS_t \sim B(1, 0.5)$. The mid-price return is drawn from a normal distribution of which the mean is zero and the variance is 4×10^{-8}, i.e. $\Delta M_t \sim N(0, 4 \times 10^{-8})$. The spread is time-varying and follows a normal distribution which mean is 0.0003 and standard deviation is 10^{-4}, i.e. $SP_t \sim N(0.03, 10^{-4})$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

3.7 Inventory Control Costs and Asymmetric Information Costs

In this section, most settings are the same as the ones in Section 3.3 except that now the mid-price return is influenced by the past trade direction, and thus there are the inventory control and the asymmetric information (IC & AS) components of the spread. Because we focus on spread estimation rather than spread decomposition, we do not distinguish the IC & AS components. Thus all the differences of the performance of the estimators can be imputed to the existence of these components. Let $\omega = 1$ in equation (26), which suggests that the mid-price is influenced by the past trade direction. The coefficient is given by $\chi = \frac{1}{3}$, which suggests that the IC and AS costs contribute one third of the total spread. In this section, trade directions are random; mid-prices are influenced by the past trade direction; and the spread is fixed. Under these circumstances, the Roll and the CS estimators are biased, and the error of the HS estimator is unbiased, because it explicitly allows for IC and AS effects. Formally, let $\varphi = 0$, $\psi = 0$, $\phi = 1$ in equation (24), which suggests that trade directions are random. The spread is fixed at 0.03. The system is given by:

$$\begin{aligned} BS_t &= 2 \cdot (\varpi_t - 0.5) \\ \varpi_t &\sim B(1, 0.5) \\ \Delta M_t &= \tfrac{1}{3}\, BS_{t-1} \cdot \tfrac{SP_t}{2} + \varepsilon_t \\ \varepsilon_t &\sim N(0, 4 \times 10^{-8}) \\ SP_t &= 0.03 \\ p_t &= M_t + \tfrac{SP_t}{2} \cdot BS_t \end{aligned} \tag{30}$$

The results are presented in Table 5, which should be compared with Table 1. The statistic ϱ reports the coefficient of an equation suggested by Stoll (1989) and represents the proportion of the IC & AS components of the spread.

The random component of mid-price returns has the same distribution as in Table 1, but now prices react positively to past trade direction, so the standard deviation of mid-price returns is greater in Table 5 than in Table 1. As predicted theoretically, the Roll estimator now underestimates substantially (by 18.3% at high frequencies), but tends to overestimate slightly at 24 hours. The HS estimator remains unbiased. The Hasbrouck estimator performs similarly to the Roll estimator with smaller standard deviations. The CS estimator underestimates considerably (by about 50%) at higher frequencies, but

overestimates by 20% at 24 hours. All three estimators produce far more variable results in individual simulations than in Table 1: for all of them, the standard deviation is much greater than in Table 1. The relative performance of the three estimators is similar to that in Table 1, with the CS estimator having the lowest RMSE at twelve and 24 hours but the highest at one hour or less. As in Table 1, estimates based on high-frequency data are much more accurate.

Table 5: Asymmetric Information and Inventory Control (Spread=0.03)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	5.00	11.2	19.4	27.4	38.8	77.6	134	190
Spread/(returns SD)	6.000	2.679	1.546	1.095	0.773	0.387	0.224	0.158
ϱ	0.333	0.333	0.333	0.333	0.328	0.312	0.247	0.200
Roll 1984								
Estimates$\times 10^{-3}$	24.5	24.5	24.5	24.5	24.4	22.6	27.6	43.1
Relative Estimate	0.817	0.817	0.817	0.817	0.813	0.753	0.920	1.437
Est-Std $\times 10^{-3}$	0.0453	0.126	0.321	0.710	1.68	13.0	27.8	46.9
RMSE$\times 10^{-3}$	5.50	5.50	5.51	5.55	5.85	14.96	27.90	48.70
Huang and Stoll 1997								
Estimates$\times 10^{-3}$	30	30	30	30	30	30.2	30.4	30.8
Relative Estimate	1.000	1.000	1.000	1.000	1.000	1.007	1.013	1.027
Est-Std$\times 10^{-3}$	0.000610	0.0684	0.220	0.446	0.886	3.56	10.5	22.0
RMSE$\times 10^{-3}$	0.000610	0.0684	0.220	0.446	0.886	3.57	10.5	22.0
Corwin and Schultz 2012*								
Estimates $\times 10^{-3}$			13.6	15.3	17.1	22.0	28.5	36.0
Relative Estimate			0.453	0.510	0.570	0.733	0.950	1.200
Est-Std $\times 10^{-3}$			0.132	0.234	0.461	1.88	5.57	9.62
RMSE$\times 10^{-3}$			16.4	14.7	12.9	8.22	5.77	11.3
Hasbrouck 2009								
Estimates $\times 10^{-3}$	25.0	24.7	24.7	24.6	24.2	21.5	38.8	59.9
Relative Estimate	0.833	0.823	0.823	0.820	0.807	0.717	1.293	1.997
Est-Std$\times 10^{-3}$	0.000726	0.0632	0.272	0.645	1.66	8.18	13.7	20.1
RMSE$\times 10^{-3}$	5.000	5.300	5.307	5.438	6.033	11.797	16.283	36.028

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is drawn from a binomial distribution, i.e. $BS_t \sim B(1, 0.5)$. The mid-price return is influenced by the past trade direction and a random shock drawn from a normal distribution of which the mean is zero and the variance is 4×10^{-8}. Thus there are the inventory control and the asymmetric information components of the spread. Formally, the mid-price returns are given by, $\Delta M_t = \frac{1}{3} BS_{t-1} \cdot \frac{SP_t}{2} + \varepsilon_t$ where $\varepsilon_t \sim N(0, 4 \times 10^{-8})$. The spread is fixed and equals to 0.03, i.e. $SP_t = 0.03$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected. ϱ reports the coefficient suggested by Stoll (1989) and represents the proportion of the inventory control and asymmetric information components of the spread.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

3.8 Negatively Auto-Correlated Mid-Price Returns

In this section, most settings are the same as the ones in Section 3.4 except that mid-price returns are autocorrelated. Thus all the differences of the performance of the estimators can be imputed to autocorrelation. Let $\tau = 1$ in equation (26), which suggests that the mid-price is autocorrelated. The coefficient is given by $\zeta = -0.3303$. Thus, the mid-price returns are negatively autocorrelated, which coincides with the results in Goodhart et al. (1996) and Daníelsson and Payne (2002), where mid-point returns of both the EFX and Reuters D2000-2 systems are shown to be negatively autocorrelated, although the autocorrelation is weaker in the Reuters D2000-2 system. In this section, trade directions are random; mid-prices are autocorrelated; and the spread is fixed at 0.0003. Under these circumstances, both the CS and Roll estimator are biased, and the error of the HS estimator is unbiased. Because the autocorrelation reduces the volatility of mid-prices, the CS estimator may have smaller errors compared to section 3.4. Formally, let $\varphi = 0$, $\psi = 0$, $\phi = 1$ in equation (24), which suggests that trade directions are random. The system is given by:

$$\begin{aligned} BS_t &= (\varpi_t - 0.5) \cdot 2 \\ \varpi_t &\sim B(1, 0.5) \\ \Delta M_t &= -0.3303\, \Delta M_{t-1} + \varepsilon_t \\ \varepsilon_t &\sim N(0, 4 \times 10^{-8}) \\ SP_t &= 0.0003 \\ p_t &= M_t + \tfrac{SP_t}{2} \cdot BS_t \end{aligned} \tag{31}$$

The results are presented in Table 6, which should be compared to Table 2 (the small spread case). The spread is the same but the standard deviation of mid-price returns is slightly larger than in Table 2, because of the autocorrelation. *CovarianceMid* reports the average covariance of mid-price returns.

The Roll estimator overestimates the spread at all time intervals, by at least 10%. This is because it attributes the entire negative autocorrelation in the observed price series to the spread, and none of it to autocorrelated mid-price returns. In the tick-by-tick case, the estimate is 3.86×10^{-4}; the covariance of the transaction price returns is 3.725×10^{-8}; and the covariance of mid-price returns is -1.48×10^{-8}. If the adjusted Roll estimator in George et al. (1991), which is true when there is autocorrelation, is used, the estimated

spread in tick-by-tick data is 0.0003. Thus, the error is 8.6×10^{-5}, which can be fully explained by the discussion in George et al. (1991).

The HS estimator remains unbiased at high frequencies. At lower frequencies it exhibits a slight upward bias, reaching nearly 10% at 24 hours, whereas in Table 2 the bias at lower frequencies was downward.

The Hasbrouck estimator performs similar to the Roll estimator with smaller errors and standard deviations.

The CS estimator exhibits a good deal of bias, as in Table 2. A feature of all three estimators is that the standard deviations of estimates from individual simulations are always a bit lower than in Table 2. Nevertheless, the RMSEs indicate that the relative performance of the estimators is similar to that in Table 2.

Table 6: Autocorrelated Mid-price Returns (Spread=0.0003)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.212	0.360	0.596	0.833	1.17	2.33	4.03	5.69
Spread/(returns SD)	1.42	0.833	0.503	0.360	0.256	0.129	0.074	0.053
CovarianceMid$\times 10^{-8}$	-1.48	-0.846	-0.834	-0.837	-0.820	-1.28	-1.57	-5.80
				Roll 1984				
Estimates $\times 10^{-3}$	0.386	0.352	0.351	0.350	0.331	0.382	0.694	1.17
Relative Estimate	1.287	1.173	1.170	1.167	1.103	1.273	2.313	3.900
Est-Std$\times 10^{-3}$	0.000856	0.00352	0.0136	0.0367	0.122	0.368	0.793	1.34
RMSE$\times 10^{-3}$	0.08600	0.0521	0.0528	0.0620	0.1259	0.377	0.885	1.598
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	0.300	0.300	0.300	0.300	0.301	0.301	0.315	0.329
Relative Estimate	1.000	1.000	1.000	1.000	1.003	1.003	1.050	1.097
Est-Std$\times 10^{-3}$	0.000632	0.00255	0.00711	0.0143	0.0282	0.108	0.328	0.673
RMSE$\times 10^{-3}$	0.000632	0.00255	0.00711	0.0143	0.0282	0.108	0.328	0.674
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			0.0734	0.128	0.185	0.333	0.528	0.704
Relative Estimate			0.245	0.427	0.617	1.110	1.760	2.347
Est-Std$\times 10^{-3}$			0.00346	0.00645	0.0133	0.0508	0.149	0.300
RMSE$\times 10^{-3}$			0.227	0.172	0.116	0.0606	0.272	0.503
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	0.336	0.344	0.348	0.348	0.303	0.481	1.09	1.74
Relative Estimate	1.120	1.147	1.160	1.160	1.010	1.603	3.633	5.800
Est-Std$\times 10^{-3}$	0.000745	0.00307	0.0132	0.0374	0.102	0.166	0.375	0.573
RMSE$\times 10^{-3}$	0.0360	0.0441	0.0498	0.061	0.102	0.246	0.874	1.550

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is drawn from a binomial distribution, i.e. $BS_t \sim B(1, 0.5)$. The mid-price return is autocorrelated and is obtained from $\Delta M_t = -0.3303\,\Delta M_{t-1} + \varepsilon_t$, where ε is a random noise of which the mean is zero and the variance is 4×10^{-8}, i.e. $\varepsilon_t \sim N(0, 4 \times 10^{-8})$. The spread is fixed and equals to 0.0003. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

Mid-price returns SD is the average of the standard deviations of mid-price returns.

CovarianceMid reports the average covariance of mid-price returns.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

3.9 Autocorrelated Trade Directions

In this section, we again use the small spreads (0.0003) of Section 3.4 , but now we allow trade directions to be positively autocorrelated. Positively autocorrelated trade directions might result, for example, from hot-potato trading. As before, mid-prices follow a random walk and the spread is fixed. Under these circumstances, the Roll estimator will underestimate the true spread, as was mentioned in theoretical analysis, but the HS estimator should remain unbiased. Formally, let $\varphi = 1$, $\psi = 0$, $\phi = 0$ in equation (24), which suggests that trade directions are autocorrelation. Let the function $F(BS_{t-1})$ make sure $Pr(BS_t = BS_{t-1}) = 0.53$. In other words, let the probability of the trade direction to be the same direction as the past one is 53%, which is slight light higher than 50%, and thus trade directions are positively autocorrelated, although the autocorrelation is not very strong. Let $\tau + \omega = 0$ in equation (26), which suggests that the mid-price follows a random walk process. The standard deviation of mid-price returns is $\sigma = 0.0002$. The system is given by:

$$\begin{aligned} & Pr(BS_t = BS_{t-1}) = 0.53 \\ & \Delta M_t = \varepsilon_t \\ & \varepsilon_t \sim N(0, 4 \times 10^{-8}) \\ & SP_t = 0.0003 \\ & p_t = M_t + \tfrac{SP_t}{2} \cdot BS_t \end{aligned} \tag{32}$$

The results are presented in Table 7, which should be compared to Table 2. The standard deviation of mid-price returns is identical to that in Table 2. The statistic δ represents the probability of trade directions continuance, which is exactly a half for time intervals of five minutes or longer, because the relatively weak autocorrelation in tick-by-tick data is quickly dissipated. If δ is known, one can apply the Choi et al. (1988) version of the Roll model to obtain the true spread. In the tick-by-tick case, the covariance of the transaction price returns is 1.97×10^{-4}, $\delta = 0.532$, and the spread, estimated by the Choi et al. (1988) version of the Roll model, is 3.00×10^{-4}. Therefore, the error can be fully explained by the discussion in Choi et al. (1988).

For all three estimators the results are similar to those shown in Table 2, except that the Roll estimator now underestimates by 6% in tick-by-tick data, because of the autocorrelation. Otherwise the relative performance of the estimators is unchanged.

Table 7: Autocorrelated trade directions (Spread=0.0003)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.200	0.447	0.774	1.09	1.55	3.09	5.36	7.58
Spread/(returns SD)	1.5	0.671	0.387	0.273	0.194	0.0968	0.0560	0.0396
δ	0.532	0.500	0.500	0.500	0.500	0.500	0.500	0.500
				Roll 1984				
Estimates $\times 10^{-3}$	0.281	0.300	0.300	0.291	0.274	0.432	0.916	1.63
Relative Estimate	0.937	1.000	1.000	0.970	0.913	1.440	3.053	5.433
Est-Std$\times 10^{-3}$	0.000910	0.00556	0.0255	0.0791	0.176	0.472	1.04	1.75
RMSE$\times 10^{-3}$	0.0190	0.00556	0.0255	0.0796	0.1779	0.490	1.209	2.198
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	0.300	0.300	0.300	0.300	0.300	0.297	0.292	0.286
Relative Estimate	1.000	1.000	1.000	1.000	1.000	0.990	0.973	0.953
Est-Std $\times 10^{-3}$	0.000606	0.00317	0.00926	0.0179	0.0362	0.143	0.448	0.888
RMSE $\times 10^{-3}$	0.000606	0.00317	0.00926	0.0179	0.0362	0.1430	0.448	0.888
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			-0.0515	0.00649	0.0718	0.256	0.494	0.741
Relative Estimate			-0.172	0.022	0.239	0.853	1.647	2.470
Est-Std $\times 10^{-3}$			0.00446	0.00868	0.0174	0.0683	0.201	0.389
RMSE $\times 10^{-3}$			0.352	0.294	0.229	0.0812	0.279	0.588
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	0.287	0.3	0.295	0.288	0.276	0.602	1.43	2.33
Relative Estimate	0.957	1.000	0.983	0.960	0.920	2.007	4.767	7.767
Est-Std$\times 10^{-3}$	0.00106	0.00526	0.0266	0.0697	0.106	0.211	0.493	0.755
RMSE$\times 10^{-3}$	0.013	0.005	0.027	0.071	0.109	0.368	1.233	2.166

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is positively autocorrelated. The probability of trade direction continuance is set to be 53%. The mid-price return is drawn from a normal distribution of which the mean is zero and the variance is 4×10^{-8}, i.e. $\Delta M_t \sim N(0, 4 \times 10^{-8})$. The spread is fixed and equals to 0.0003, i.e. $SP_t = 0.0003$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

Mid-price returns SD is the average of the standard deviations of mid-price returns.

δ is the probability of the trade direction keeping the same direction as the past one.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

3.10 Feedback Trading

In this section, we consider the influence of an alternative trade-direction-generating process, by introducing (positive) feedback training. Thus all the differences in the performance of the estimators can be imputed to feedback trading. In this section, trade directions are positively related to past mid-price returns (i.e. a buy order for a currency is more likely after its price has risen); mid-prices follow a random walk process; the volatility of mid-price returns is small; and the spread is fixed at 0.0003. Under these circumstances, both the HS and the Roll estimator are biased, and the error of the CS estimator should also be influenced by feedback trading. Formally, let $\varphi = 0$, $\psi = 1$, $\phi = 0$ in equation (24), which suggests that trade directions are a function of past mid-price returns. Let $\kappa = 0.65$, which suggest there is positive feedback trading. Let $\tau + \omega = 0$ in equation (26), which suggests that the mid-price follows a random walk process. The system is given by:

$$
\begin{aligned}
& I_t(\Delta M) \sim \begin{cases} B(1, 0.65) & if \quad \Delta M_t > 0 \\ B(1, 0.35) & if \quad \Delta M_t < 0 \end{cases} \\
& BS_t = (I_t - 0.5) \cdot 2 \\
& \Delta M_t = \varepsilon_t \\
& \varepsilon_t \sim N(0, 9 \times 10^{-12}) \\
& SP_t = 0.0003 \\
& p_t = M_t + \tfrac{SP_t}{2} \cdot BS_t
\end{aligned}
\tag{33}
$$

The results are presented in Table 8, which should be compared with Table 2. $Cov(\Delta M_t, BS_t)$ represents the covariance of mid-price returns and trade directions, which suggests the existence of feedback trading.

The Roll estimator overestimates the true spread. Even in tick-by-tick data, the bias is 15%. Its standard deviation is slightly higher than in Table 2, but not materially so. The Hasbrouck estimator is similar to the Roll estimator but has smaller standard deviations in longer time intervals.

The HS estimator is badly affected by feedback trading, overestimating by 32% even in tick-by-tick data. Like the Roll and Hasbrouck estimators, its standard deviation is little affected. Because of the larger bias, the HS estimator has a higher RMSE than the Roll and Hasbrouck estimators at short time intervals (the exception is four hours). This is

the only case, of those that we have examined, where the Roll and Hasbrouck estimators outperform the HS estimator.

The CS estimator tends to produce higher estimates in all time intervals than those in Table 2. As in Table 2, it underestimates badly at high frequencies. At lower frequencies it has a lower RMSE than the Roll estimator, as in Table 2, because the standard deviation does not increase so sharply with the time interval.

3.11 Summary

Where high-frequency data are available, the spread is best estimated on the highest possible sampling frequency by the HS method, if trade direction is known, or by the Roll or Hasbrouck estimator if not. Low-frequency estimators like CS discard too much information to be competitive. However, when only low-frequency (e.g. daily) data are available, the CS estimator generally outperforms the others. A notable exception and it is an important one in practice is that the CS estimator seriously overestimates the spread when it varies over time (e.g. across trading hours or days of the week), to the extent that it is inferior to the other estimators in these circumstances, even with low-frequency sampling. The larger the sample, the better the relative performance of the HS estimator in low-frequency data; indeed if the sample is large enough it outperforms the CS estimator.

Tables 9 and 10 show the ranking of the estimators under various conditions according to the RMSE. In most high-frequency cases (Table 9) except for the case of feedback trading, the HS estimator is preferred and the Hasbrouck estimator is the second. In most low-frequency cases (Table 10) the CS estimator is the first choice and the HS estimator is on the second place. Furthermore, according to Table 3, when the sample size is large, the HS estimator is preferred.

Table 8: Feedback Trading (Spread=0.0003)

	Tick	5-Min	15-Min	30-Min	1-Hour	4-Hour	12-Hour	24-Hour
Mid-price returns SD $\times 10^{-3}$	0.200	0.447	0.774	1.09	1.55	3.09	5.36	7.58
Spread/(returns SD)	1.5	0.671	0.387	0.273	0.194	0.0968	0.0560	0.0396
$Cov(\Delta M_t, BS_t)$ $\times 10^{-3}$	0.0479	0.0479	0.0479	0.0480	0.0477	0.0447	0.0584	0.0544
				Roll 1984				
Estimates $\times 10^{-3}$	0.345	0.345	0.344	0.336	0.307	0.450	0.923	1.57
Relative Estimate	1.150	1.150	1.147	1.120	1.023	1.500	3.077	5.233
Est-Std$\times 10^{-3}$	0.000925	0.00523	0.0219	0.0661	0.174	0.489	1.07	1.78
RMSE$\times 10^{-3}$	0.0450	0.0453	0.0492	0.0753	0.174	0.511	1.238	2.187
				Huang and Stoll 1997				
Estimates $\times 10^{-3}$	0.396	0.396	0.396	0.396	0.398	0.406	0.412	0.431
Relative Estimate	1.320	1.320	1.320	1.320	1.327	1.353	1.373	1.437
Est-Std$\times 10^{-3}$	0.000587	0.00302	0.00919	0.0179	0.0373	0.143	0.437	0.862
RMSE$\times 10^{-3}$	0.0960	0.0961	0.0964	0.0977	0.105	0.178	0.451	0.872
				Corwin and Schultz 2012*				
Estimates $\times 10^{-3}$			-0.0135	0.0460	0.111	0.298	0.545	0.770
Relative Estimate $\times 10^{-3}$			-0.045	0.153	0.370	0.993	1.817	2.567
Est-Std$\times 10^{-3}$			0.00451	0.00869	0.0169	0.0678	0.212	0.407
RMSE$\times 10^{-3}$			0.314	0.254	0.190	0.0678	0.324	0.622
				Hasbrouck 2009				
Estimates $\times 10^{-3}$	0.351	0.346	0.341	0.331	0.296	0.617	1.45	2.32
Relative Estimate	1.170	1.153	1.137	1.103	0.987	2.057	4.833	7.733
Est-Std$\times 10^{-3}$	0.000862	0.0049	0.0224	0.0666	0.113	0.223	0.532	0.783
RMSE$\times 10^{-3}$	0.051	0.046	0.047	0.073	0.113	0.388	1.267	2.166

There are 1000 replications. There are 432000 periods, each of which represents one minute, in each replication. Data of each replication are generated according to the following system. The trade direction is positively autocorrelated. The mid-price return is drawn from a normal distribution of which the mean is zero and the variance is 4×10^{-8}, i.e. $\Delta M_t \sim N(0, 4 \times 10^{-8})$. trade directions is positively correlated to mid-price returns. The probability of a buy (sell) order being after a positive (negative) return is 65%. i.e. The spread is fixed and equals to 0.0003, i.e. $BS_t \sim B(1, 0.65)$ if $\Delta M_t > 0$ and $BS_t \sim B(1, 0.35)$ if $\Delta M_t < 0$. $SP_t = 0.0003$. The transaction price is the mid-price plus or minus a half spread, i.e. $p_t = M_t + \frac{SP_t}{2} \cdot BS_t$. Each replication is also sampled at longer time intervals: five-minute, fifteen-minute, thirty-minute, one-hour, four-hour, twelve-hour and twenty-four- hour, and only the last observation is kept. Thus, there are eight subgroups for each replication. For each subgroup, the standard deviation of mid-price returns, and the estimated spread are collected.

$Cov(\Delta M_t, BS_t)$ is the covariance of mid-price returns and trade directions, which reflects the existence of feedback trading.

* The CS estimate can be obtained either by taking an average of the estimates within the replication or by calculating the averages of β and γ within the replication. Because the second method performs worse than the first one in most cases, we do not report the results of the second method.

The other settings are the same as Table 1

Table 9: The ranking of the estimators by RMSE: high sampling frequencies

	Roll	Huang & Stoll	Corwin & Schultz	Hasbrouck
Fixed Big Spread	2	1	4	3
Fixed Small Spread	2	1	4	3
Time-varying Spreads	3	1	4	2
ASIC	3	1	4	2
Autocorrelated Mid-price Return	3	1	4	2
Autocorrelated trade directions	3	1	4	2
Feedback Trading	1	3	4	2

This table summaries the main findings of the simulation experiments in Section 3. The estimator with the lowest RMSE is ranked "1", etc.

Table 10: The ranking of the estimators by RMSE: low sampling frequencies

	Roll	Huang & Stoll	Corwin & Schultz	Hasbrouck
Fixed Big Spread	4	3	1	2
Fixed Small Spread	4	2	1	3
Time-varying Spreads	3	1	4	2
ASIC	4	2	1	3
Autocorrelated Mid-price Return	4	2	1	3
Autocorrelated trade directions	4	2	1	3
Feedback Trading	4	2	1	3

This table summaries the main findings of the simulation experiments in Section 3. The estimator with the lowest RMSE is ranked "1", etc.

4 Conclusion

The appropriate spread estimator to use is likely to vary with the characteristics of the data and the frequency with which it is sampled. Most estimators perform well when conditions are ideal (the spread is large and the data conform to the assumptions underlying the estimator). More interesting are the cases where these assumptions break down. We have performed simulations assuming one trade per minute in a continuously open market, with sampling frequencies varying from one minute to 24 hours. We have considered the Huang-Stoll (1997) estimator, which requires closing prices and trade directions; the Roll (1984) and and Hasbrouck (2004, 2009) estimators (closing prices only); and the Corwin-Schultz (2010) estimator, which uses only the highest and the lowest price recorded in an interval, and is explicitly designed for low-frequency (e.g. daily) data.

The first result is that the best estimates come from using the highest frequency of data available. The main reason is that, the longer the sampling interval, the greater is the price volatility over the interval, and therefore the harder it is to estimate the spread, since price volatility represents the "noise" that the estimator is trying to distinguish from the "signal". All estimators therefore lose accuracy as the sampling interval increases.

The second finding is that the relative performance of estimators can vary quite markedly with the sampling frequency. For example, the Corwin-Schultz estimator performs relatively poorly at small sampling intervals (less than one hour). At high frequencies it is always inferior to the Roll and Hasbrouck estimators, which also does not require trade direction data. But the Corwin-Schultz estimator loses accuracy as the sampling frequency increases less fast than the other estimators, and this means that at lower frequencies (four hours or more), it often has the lowest root mean square error.

Our conclusions are different according to whether or not high-frequency data are available. If they are, the optimal procedure is to use the Huang-Stoll estimator (if trade directions are known), or Roll-type estimators if they are not. The one exception to this is when there is feedback trading. In the presence of feedback trading, the Roll and Hasbrouck estimators outperform the Huang-Stoll estimator. When only low-frequency data are available, all estimators are inaccurate, but the Corwin-Schultz estimator tends to be the least inaccurate. There is one exception to this: when the spread is time-varying, the Corwin-Schultz estimator is seriously upwardly biased, and it is better to use

the Huang-Stoll estimator (if possible) or the Hasbrouck estimator.

Another issue that influences the performance of estimators is the sample size. Increasing the sample size will significantly reduce the standard deviation of estimates. When the sample size is large, the estimator with small error, such as the Huang-Stoll estimator, is preferred because the standard deviation is not the dominate factor.

A natural question to ask is what the effect would be if two or more of the conditions that we investigate occurred together. We have conducted experiments of this kind (not reported here), and we have found that the effects are generally additive, i.e. the bias associated with conditions A and B together tends to be similar to the sum of the biases associated with A and B individually.

Appendix

Error one of the CS estimator

According to the assumption of the CS estimator, the relationship between the highest (lowest) observed prices and the highest (lowest) mid-prices is given by:

$$H_t^O = TH_t^M + \frac{\widehat{SP}}{2}; \qquad L_t^O = TL_t^M - \frac{\widehat{SP}}{2} \tag{34}$$

where TH^M (TL^M) is the true highest (lowest) mid-price and $\widehat{SP}$ is the estimated spread. Then $\widehat{SP}$ is given by:

$$(H_t^O - L_t^O) - (TH_t^M - TL_t^M) = \widehat{SP} \tag{35}$$

In fact, the reality is that:

$$H_t^O = H_t^M + \frac{SP}{2}; \qquad L_t^O = L_t^M - \frac{SP}{2} \tag{36}$$

where

$$TH_t^M \geq H_t^M; \qquad TL_t^M \leq L_t^M \tag{37}$$

The inequalities suggest that in reality, the assumption is not always true, in other words, not all the highest (lowest) prices are used. If the assumption were correct, which suggests that $TH_t^M = H_t^M$ and $TL_t^M = L_t^M$, then the spread is correctly estimated. If the assumption were not true, which suggests that $TH_t^M > H_t^M$ or $TL_t^M < L_t^M$, then the spread is underestimated. The error is given by:

$$e_m = SP - \widehat{SP} = D_H + D_L \tag{38}$$

where the distances D_H and D_L are given by,

$$\begin{aligned} D_H &= TH_t^M - H_t^M \\ D_L &= L_t^M - TL_t^M \end{aligned} \tag{39}$$

The probability of the events that $TH_t^M > H_t^M$ or $TL_t^M < L_t^M$ happen is positively correlated with the mid-price volatility and is negatively related to the number of observations in an interval. It is not a problem if observations in the interval are infinite, because the distance should be very close to zero. Given the fact that the CS estimator usually works

on daily data, the number of observations in one interval is limited, the distance is positively correlated with the volatility of mid-price returns. If the number of observations is fixed in an interval, a bigger volatility means a larger diffusion, and thus the average distance between two observations is larger. To sum up, the greater the volatility the bigger the error.

Error three of the CS estimator

While Parkinson's (1980) volatility estimator works when the sample size is large, equation (16) applies it using one observation (a high-low ratio over two periods) on the left-hand side and applies it using two observations (summation of two high-low ratios of two periods) on the right hand side. It is reasonable that two-period estimation is more accurate than one-period estimation. The imbalance of the accuracy brings errors. The CS estimator assumes that the relationship of the high-low ratios is valid for each pair of two adjacent periods,

$$\begin{aligned}\left(TH^M_{t,t+1} - TL^M_{t,t+1}\right) &= \sqrt{2} \cdot \left[\sum_{J=0}^{1}\left(TH^M_{t+J} - TL^M_{t+J}\right)\right] \\ \left(TH^M_{t,t+1} - TL^M_{t,t+1}\right)^2 &= \sum_{J=0}^{1}\left(TH^M_{t+J} - TL^M_{t+J}\right)^2\end{aligned} \tag{40}$$

In fact, the equations above are valid only in the case of expectations of them were taken. Therefore, the true version of equation (14) is:

$$\left(H^O_{t,t+1} - L^O_{t,t+1}\right)^2 + \epsilon_1 = \left(TH^M_{t,t+1} - TL^M_{t,t+1}\right)^2 + \epsilon_3 + 2SP \cdot \left[\left(TH^M_{t,t+1} - TL^M_{t,t+1}\right) + \epsilon_2\right] + SP^2 \tag{41}$$

Since ϵ_1, ϵ_2 and ϵ_3 are shocks corresponding to a same high-low ratio of one-period estimation, and thus are highly correlated, they can be partially cancelled out in equation (14). Formally,

$$e_{v1} = \epsilon_1 - 2SP \cdot \epsilon_2 - \epsilon_3 \approx 0 \tag{42}$$

Equations (41) are two-period versions of equation (15), therefore, e_{v1} shares the same intuition as the error in the first equation of the estimator system (i.e. e_β).

The CS estimator links the high-low ratio over two periods and two high-low ratios of the single periods. Similarly, the assumption of the link is only valid when the expectations

are taken. The true version of equation (14) is:

$$\begin{aligned}\left(H^{O}_{t,t+1} - L^{O}_{t,t+1}\right)^2 + \epsilon_1 = &\left[\sum_{J=0}^{1}\left(TH^{M}_{t+J} - TL^{M}_{t+J}\right)^2 + \epsilon_5\right] + 2SP \\ &\cdot\sqrt{2}\left[\sum_{J=0}^{1}\left(TH^{M}_{t+J} - TL^{M}_{t+J}\right) + \epsilon_4\right] + SP^2\end{aligned} \tag{43}$$

In contrast to equations (41), ϵ_4 and ϵ_5 are with two-period estimation, the correlation between ϵ_1, ϵ_4 and ϵ_5 is not as high as the one with ϵ_2 and ϵ_3. Thus,

$$e_{v2} = \epsilon_1 - 2SP \cdot \epsilon_4 - \epsilon_5 \neq 0 \tag{44}$$

e_{v2} is the error of equation (16) (e_γ). The expectations of both e_{v1} and e_{v2} are zeros. The variances could be more important in a non-linear system. The variance of e_{v2} is much greater than that of e_{v1} and is less than $3k_1\sigma^2$ but is still positively correlated with the volatility of mid-price returns.

References

Amihud, Y. and H. Mendelson (1980). Dealership market : Market-making with inventory. *Journal of Financial Economics 8*(1), 31–53.

Anand, A. and A. K. Karagozoglu (2006). Relative performance of bidask spread estimators: Futures market evidence. *Journal of International Financial Markets, Institutions and Money 16*(3), 231 – 245.

ap Gwilym, O. and S. Thomas (2002). An empirical comparison of quoted and implied bid-ask spreads on futures contracts. *Journal of International Financial Markets, Institutions and Money 12*(1), 81–99.

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.

Berger, D. W., A. P. Chaboud, S. V. Chernenko, E. Howorka, and J. H. Wright (2008). Order flow and exchange rate dynamics in electronic brokerage system data. *Journal of International Economics 75*(1), 93–109.

Bessembinder, H. (1994). Bid-ask spreads in the interbank foreign exchange markets. *Journal of Financial Economics 35*(3), 317–348.

Chan, K. C., W. G. Christie, and P. H. Schultz (1995). Market structure and the intraday pattern of bid-ask spreads for nasdaq securities. *The Journal of Business 68*(1), 35–60.

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.

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

Daníelsson, J. and R. Love (2006). Feedback trading. *International Journal of Finance & Economics 11*(1), 35–53.

Daníelsson, J. and R. Payne (2002). Real trading patterns and prices in spot foreign exchange markets. *Journal of International Money and Finance 21*(2), 203–222.

De Long, J. B., A. Shleifer, L. H. Summers, and R. J. Waldmann (1990). Positive feedback investment strategies and destabilizing rational speculation. *The Journal of Finance 45*(2), 379–395.

Evans, M. D. D. and R. K. Lyons (2002). Order flow and exchange rate dynamics. *The Journal of Political Economy 110*(1), 170–180.

Garman, M. B. and M. J. Klass (1980). On the estimation of security price volatilities from historical data. *The Journal of Business 53*(1), 67–78.

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

Goodhart, C., T. Ito, and R. Payne (1996). One day in june 1993: A study of the working of reuters d2000-2 electronic foreign exchange trading system. In J. Frankel, G. Galli, and A. Giovannini (Eds.), *The Microstrttcture of Foreign Exchange Markets*. Chicago: University of Chicago Press.

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. (1991). Measuring the information content of stock trades. *The Journal of Finance 46*(1), 179–207.

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.

Ho, T. S. Y. and H. R. Stoll (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics 9*(1), 47 – 73.

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

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.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica 53*(6), 1315–1335.

Lin, C.-C. (2013). Estimation accuracy of high-low spread estimator. *Finance Research Letters*, forthcoming.

Lyons, R. K. (1995). Tests of microstructural hypotheses in the foreign exchange market. *Journal of Financial Economics 39*(2-3), 321–351.

Lyons, R. K. (1997). A simultaneous trade model of the foreign exchange hot potato. *Journal of International Economics 42*(3-4), 275–298.

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.

McInish, T. H. and R. A. Wood (1992). An analysis of intraday patterns in bid/ask spreads for nyse stocks. *The Journal of Finance 47*(2), 753–764.

Moulton, P. C. (2005). You cant always get what you want: Trade-size clustering and quantity choice in liquidity. *Journal of Financial Economics 78*(1), 89 – 119.

Nofsinger, J. R. and R. W. Sias (1999). Herding and feedback trading by institutional and individual investors. *The Journal of Finance 54*(6), 2263–2295.

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.

Sias, R. W. (2004). Institutional herding. *The Review of Financial Studies 17*(1), 165–206.

Sias, R. W. and L. T. Starks (1997). Return autocorrelation and institutional investors. *Journal of Financial Economics 46*(1), 103–131.

Stoll, H. R. (1978). The supply of dealer services in securities markets. *The Journal of Finance 33*(4), pp. 1133–1151.

Stoll, H. R. (1989). Inferring the components of the bid-ask spread: Theory and empirical tests. *The Journal of Finance 44*(1), 115–134.