NCCC-134

APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT

# Measuring Liquidity Costs in Agricultural Futures Markets

by

Julieta Frank and Philip Garcia

# Measuring Liquidity Costs in Agricultural Futures Markets

***Julieta Frank***

**and**

***Philip Garcia****

*Paper presented at the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management*
*Chicago, Illinois, April 16-17, 2007*

* Julieta Frank is a Graduate Research Assistant (jfrank1@uiuc.edu) and Philip Garcia is the T. A. Hieronymus Distinguished Chair in Futures Markets in the Department of Agricultural and Consumer Economics at the University of Illinois at Urbana-Champaign. This work was partially supported by the National Center for Supercomputing Applications under grant number SES060005N and utilized the IBM P690 (Teragrid).

# Measuring Liquidity Costs in Agricultural Futures Markets

*Estimation of liquidity costs in agricultural futures markets is challenging because bid-ask spreads are usually not observed. Spread estimators that use transaction data are available, but little agreement exists on their relative accuracy and performance. We evaluate four conventional and a recently proposed Bayesian estimators using simulated data based on Roll's standard liquidity cost model. The Bayesian estimator tracks Roll's model relatively well except when the level of noise in the market is large. We derive an improved estimator that seems to have a higher performance even under high levels of noise which is common in agricultural futures markets. We also compute liquidity costs using data for hogs and cattle futures contracts trading on the Chicago Mercantile Exchange. The results obtained for market data are in line with the findings using simulated data.*

**Keywords**: liquidity costs, bid-ask spread, Bayesian estimation, Gibbs sampler

## Introduction

The cost of liquidity, often referred to as the bid-ask spread, is the difference between the prices for immediate purchase and sale (Bryant and Haigh 2004). This component of the transaction costs is usually ignored in analyses of futures markets as bids and offers occur in an open outcry pit and are not recorded. To circumvent this problem, spread estimators have been proposed that use transaction data only. Some examples are serial covariance estimators (Roll 1984; Choi, Salandro and Shastri 1988; Chu, Ding, and Pyun 1996), and mean absolute price change estimators (Thompson and Waller 1988; Wang et al. 1997). However, research suggests that these estimators are biased with respect to actual bid-ask spreads, with little agreement on the direction and magnitude of the bias. For example, Locke and Venkatesh (1997) and Ferguson and Mann (2001) show that both serial covariance estimators and mean absolute price change estimators are upward biased using audit data.[1] In contrast, Bryant and Haigh (2004) and Anand and Karagozoglu (2006) find that both serial covariance and mean absolute price change estimators are downward biased with respect to bid-ask quotes.[2]

Conventional spread estimators have been commonly used because of both the lack of alternative estimators and their simplicity. However, previous research has shown their weaknesses. For example, the covariance between adjacent price changes can yield positive values when using serial covariance estimators, making it difficult to obtain spread estimates. The Thompson-Whaley estimator can fail to distinguish between true price change volatility and volatility attributable to the bid-ask price bounce (Smith and Whaley 1994). Locke and Venkatesh (1997) and Smith and Whaley (1994) suggest that Roll's estimator is inadequate for futures markets. Chu, Ding and Pyun (1996) find that, in general, Roll's measure overestimates the bid-ask spread in foreign exchange futures prices because it fails to account for the possibility that buys and sells might not be equally likely. The above research suggests that further investigation of conventional spread estimators as well as alternative measures is needed.

Recently, Hasbrouck (2004) implemented a Bayesian Markov Chain Monte Carlo (MCMC) algorithm, the Gibbs sampler, to generate estimates of liquidity costs. Bayesian techniques are attractive in this context for a number of reasons. Estimation is based on parameters' posteriors which incorporate all the information in the observed transaction prices. Like the mean absolute price change estimators, this procedure does not contain the problem of unfeasible values (i.e. positive covariance between adjacent price changes) because the parameters are random draws from their conditional distributions. In addition, unobserved latent variables, like the trade direction indicator, are estimated conditional on observed transaction prices rather than derived from tick rules. However, the Bayesian estimator has been less explored and its level of accuracy is not well documented. Moreover, Hasbrouck's estimates for pork bellies are considerable smaller relative to Roll's conventional estimates.

Accurate estimates of liquidity costs are of interest to exchanges, market participants, and researchers. For exchanges, knowing the cost of providing liquidity in their different markets can help develop strategies for the development of new products as well as in the regulation of market-making services of existing products. For example, liquidity costs might be useful to assess the quality of the hedging service provided by a futures contract (Pennings and Meulenberg 1997). For market participants, estimates of liquidity costs in different markets and exchanges are useful in making operational decisions. Brorsen, Buck, and Koontz (1998) suggest that wheat hedgers would maximize their utility by choosing the Chicago Board of Trade (CBOT) if they are slightly risk averse and face high liquidity cost differences, but the Kansas City Board of Trade (KCBT) is a better (utility maximizing) option if they are faced with low liquidity cost differences. For researchers, understanding the structure of liquidity costs in futures markets may provide a more comprehensive view of the pricing process. For example, Phillips and Smith (1980) show that failure to account for bid-ask spreads in the calculation of excess returns in options and stock markets leads to abnormal returns and to the wrong conclusion of market inefficiency. Much research has been done in stock markets, however, futures markets have been less explored due to the lack of bid-ask quotes.

The purposes of the research are to evaluate commonly used and alternative (i.e., Bayesian) spread estimators, determine their source of bias, and identify the most adequate measure of liquidity costs for different market conditions when the only data available are transaction prices. The analysis focuses on the Bayesian estimator as there are no previous studies in this arena. Due to the lack of actual bid-ask spreads, we assess the accuracy of the different measures using simulated data based on the Roll model of spread behavior. We use the mean square error and correlations to assess the performance of the spread estimators. We also compare different spread estimates using data from the lean hogs and live cattle contracts trading on the CME.

## Literature Review

Considerable research has been performed on market microstructure in general, and on liquidity in particular for stock and financial futures markets. Studies on commodity futures markets, however, are more scarce. Moreover, as argued by Bryant and Haigh (2004) and Ferguson and Mann (2001), findings from financial markets are not always directly applicable to commodity markets because both markets have different levels of transparency (i.e. amount of information available to market makers and other participants). Hence, our discussion focuses on research performed in commodity futures markets.

Research in liquidity costs for agricultural futures markets can be classified in those studies addressing specific characteristics of the different markets and those dealing with measurement problems. Very few studies exist in the first group (Thompson and Waller 1987 and 1988, and Thompson, Eales, and Seibold 1993). In the second group most of the studies stress the problems associated with spread estimators, specially when the Roll and Thompsom-Waller measures are used (for example, Ma, Peterson, and Sears 1992, Smith and Whaley 1994, and Bryant and Haigh 2004).

Thompson and Waller (1987) studied coffee and cocoa contracts in the New York Board of Trade (NYBOT) over the three-year period 1981-83. Under the hypothesis of negative price correlation, they use the average absolute value of price changes to measure execution costs. Negative price correlation emerges because market makers fill buy orders at a higher price than sell orders. Their findings show lower execution costs in actively traded nearby contracts relative to thinly traded more distant contracts.

Thompson and Waller (1988) analyzed liquidity costs for corn and oats contracts traded in the Chicago Board of Trade (CBOT) in 1984 and 1986. Their results are mixed and in some cases not consistent with expectations when Roll's serial covariance measure is used. For example, they found that the trading volume is negatively related to liquidity costs when they use the absolute price change as a proxy but has a positive relationship when they use Roll's measure as a proxy. Based on theory and past research, an increase in trading volume is expected to be negatively related to liquidity costs because it is associated with a faster rate of inventory turnover thus reducing the time and information risk faced by the market maker associated with a change in real price.

Thompson, Eales, and Seibold (1993) compared liquidity costs for the same commodity traded in different exchanges, i.e. wheat in the CBOT and in the KCBT in 1985. Using Roll's measure and the average absolute price changes measure, their results suggest that in Kansas City liquidity costs are significantly higher due, to some extent, to its lower trading volume. In both exchanges, liquidity costs are higher and more sensitive to trading volume at expiration.

Ma, Peterson, and Sears (1992) investigate the intraday behavior of selected futures contracts, including corn and soybeans. They argue that the Thompson –Waller estimator might be upward biased, and that the Roll estimator might overstate the actual spread when transaction prices are recorded on a tick basis. However, all three estimators basically reflect the U-shape price behavior during the day which suggests that different spread estimators are correlated as they capture spread variability during the day.

Smith and Whaley (1994) recognize the problems associated with the Roll and Thompson-Waller estimators. They point out that the Roll estimator becomes troublesome when the covariance between adjacent price changes is positive, and that the Thompson-Waller estimate gives an upward bias of the spread because it fails to recognize the variance of true price changes contained in the absolute value of price changes. To overcome this problem, Smith and Whaley suggest a new spread estimator based on the first two moments of absolute price change distribution. Their estimator is derived for tick-basis datasets, and is robust to different levels of serial correlation and volatility of true price changes when both simulated data and S&P500 futures data for the period 1982-1987 are used. However, their model assumes that true price changes are distributed normally with mean zero and variance $\sigma^2$ which not always holds.

Locke and Venkatesh (1997) compute futures transactions costs for several commodities to assess the performance of commonly used spread estimators. Transaction costs are defined as dollar flows from customers to market-makers, and are estimated as the difference between the average purchase price and the average sale price for all futures customers, with prices weighted by transaction size. Transactions costs are measured directly using data provided by the CFTC which contains information about each trade (commodity, delivery month, quantity, price, date and time), the trade direction (whether it is a buy or a sell), and the executing trader (whether it is a customer or a market maker). Roll's (1984) and Smith and Whaley's (1994) estimates for live cattle and lean hogs in the CME appear to be higher than the observed transaction costs (all but Roll's estimate for lean hogs are significantly higher at the 5% level).

Ferguson and Mann (2001) use a similar approach as Locke and Venkatesh (1997) to study agricultural commodities trading in the CME, namely live cattle, pork bellies, hogs, feeder cattle and lumber. Their results are consistent with Locke and Venkatesh as the observed execution spreads are lower than estimated spreads using Roll's serial covariance measure. For example, the article finds execution spreads for live cattle and lean hogs that are 85% and 73% lower than their respective estimates using Roll's measure.

Bryant and Haigh (2004) contrast observed and estimated spreads in commodity futures markets. Observed bid and ask prices, as well as transaction prices, are taken from the LIFFE for cocoa and coffee. Estimated spreads are computed using serial covariance and absolute price change measures. Observed spreads are higher than estimated spreads for both serial covariance and absolute price change measures. The correlation between estimated and observed spreads is higher for the absolute price change measures than for the serial covariance measures. Also, absolute price change estimators perform better than serial covariance estimators when evaluated using the bias and the mean square error criteria, however the latter show lower error variances. Similar results regarding the downward bias in spread estimators are found by Anand and Karagozoglu (2006) when estimated spreads are contrasted with actual spreads from the SFE in financial futures markets. These findings imply that spread estimators might not be reliable and alternative measures of liquidity costs are needed when observed bid and ask prices are not available, as is the case for major US exchanges.

## Methods

### *The model*

We use Roll's standard model as a framework to generate the simulated data which will be used to assess the performance of the other procedures. The basic Roll model is commonly used in the microstructure literature to relate transaction prices and liquidity costs. It is also the underlying model assumed by spread estimators as shown below. In the model, in the absence of transaction costs the efficient price $m_t$ reflects all available public information and follows a random walk. Futures markets operate through dealers who offer bid ($B_t$) and ask ($A_t$) prices, so that buyers buy at the price $A_t$, sellers receive the price $B_t$, and the cost of a transaction is $c$. Prices are affected by the direction of trade, $q_t=\{-1$ for a sell, $+1$ for a buy$\}$. When trading is observed, the transaction prices $p_t$ are determined using (1) and (2).

$$m_t = m_{t-1} + u_t \qquad u_t \sim iid\ N(0, \sigma^2_u) \qquad (1)$$
$$p_t = m_t + cq_t \qquad q_t \sim Bernoulli(1/2) \qquad (2)$$

When $q_t = -1$ then $p_t = B_t$ and when $q_t = +1$ then $p_t = A_t$, with $c$ being a measure of the half spread.

### *Conventional estimators*

Conventional spread measures that use transaction prices only are based on price changes negative serial dependence and prices absolute price changes. Serial covariance estimators have been proposed by Roll (1984), Choi, Salandro and Shastri (1988), and Chu, Ding and Pyun (1996). Roll is the most extensively used estimator while the other two are used for market conditions associated with different distributions of $q_t$. Since Choi, Salandro and Shastri is a special case of Chu, Ding and Pyun, it is not included in the analysis. Mean absolute price change estimators include Thompson and Waller (1987) and the CFTC measure described in Wang, Moriarty, Michalski and Jordan (1990) and Wang, Yau and Baptiste (1997).

Roll (1984) argues that trading costs induce negative serial dependence in successive observed price changes and derives the first serial covariance estimator. The assumptions are: 1) the market is informationally efficient, 2) observed price changes follow a stationary probability distribution, 3) all trades are made through the market maker who maintains a constant spread, 4) the direction of the trade is independent of the efficient price movement, i.e. $q_t$ is independent of $\Delta m_t = u_t$, and 5) each transaction is equally likely to be a purchase or a sale, i.e. $q_t \sim Bernoulli(1/2)$ in (2). Under these assumptions and taking the covariance of subsequent price changes yields,

$$\Delta p_t = u_t + c\Delta q_t \qquad u_t \sim iid\ N(0, \sigma^2_u) \qquad (3)$$
$$\text{Cov}\ (\Delta p_t,\Delta p_{t-1}) = E[u_t\ u_{t-1}] + c(E[u_t\Delta q_{t-1}] + E[u_{t-1}\Delta q_t]) + c^2\Delta q_t\Delta q_{t-1} = - c^2 \qquad (4)$$

The first term on the RHS of equation (4), $E[u_t\ u_{t-1}]$, vanishes under market efficiency as there is no information from $t-1$ contained in $t$. The second term, $E[u_t\Delta q_{t-1}] + E[u_{t-1}\Delta q_t]$, is also zero as the fourth assumption states that the direction of incoming orders does not provide information that is reflected in the efficient price. That is, changes in the direction of trades are non-informative. Solving for the last term and rearranging yields the Roll estimator for the half spread, *c*,

$$RM = \sqrt{-\operatorname{cov}(\Delta p_t, \Delta p_{t-1})} \tag{4.1}$$

where $\Delta p_t$, $t = 1,\ldots,T$, are the observed transaction prices in first differences.

Choi, Salandro and Shastri (1988) relax the assumption that each transaction is equally likely to be a purchase or a sale, i.e. $q_t \sim Bernoulli(1/2)$ in (2), and incorporate the possibility of serial correlation in transactions. Serial correlation in transactions might be due to floor brokers executing market orders for a large number of shares and splitting them among quotations of other market participants. An observer would then see successive transactions of the same type. Serial correlation may also emerge when limit orders get executed at the same time after a certain price change. Chu, Ding, and Pyun (1996) extend this idea by incorporating a longer memory in the model. In addition to the one-period conditional probabilities, $\delta = P(A_t \mid A_{t-1}) = P(B_t \mid B_{t-1})$, they define two-period conditional probabilities, $\alpha = P(A_{t+1} \mid B_{t-1}A_t) = P(B_{t+1} \mid A_{t-1}B_t)$,

$$CDP = \frac{1}{2}\sqrt{\frac{-\operatorname{cov}(\Delta p_t, \Delta p_{t+1})}{(1-\delta)(1-\alpha)}} \tag{5}$$

$$\hat{\delta} = \frac{1}{n}\sum_{t=1}^{n} T_i$$

$$\hat{\alpha} = \frac{N(a) + N(c)}{N(a) + N(1-a) + N(c) + N(1-c)}\,. \tag{6}$$

Here $\hat{\delta}$ is the maximum likelihood estimator of $\delta$, $n$ is the number of transactions, $T_i = 1$ if transaction $i$ is the same type as $i$-1 and 0 otherwise, $\hat{\alpha}$ is the maximum likelihood estimator of $\alpha$, and $N(\bullet)$ is the number of observations corresponding to each of the following events: $a = P(A_{t+1} \mid B_{t-1}A_t)$, $c = P(B_{t+1} \mid A_{t-1}B_t)$, $1\text{-}a = P(B_{t+1} \mid B_{t-1}A_t)$, $1\text{-}c = P(A_{t+1} \mid A_{t-1}B_t)$. Transactions are classified as bid and asks following Lee and Ready's (1991) tick rule. A transaction is an ask if it occurs at an uptick or zero-uptick, and as a bid if it occurs at a downtick or zero-downtick. A trade is an uptick (downtick) if the price is higher (lower) than that of the previous trade. When a trade occurs at the same price as the previous trade's, it is a zero-uptick (zero-downtick) if the last price change was an uptick (downtick) or zero-uptick (zero-downtick). The *CDP* is a more generalized estimator than *RM*; if $\alpha = \delta = 0.5$, *CDP* reduces to *RM*.

Thompson and Waller (1987) propose a measure of liquidity costs that is also based on the negative dependence of price changes. They argue that the placement of a buy order is likely to increase the average price level and the placement of a sell order is likely to decrease the average price level. When this is true for all orders, the mean absolute value price change is an unbiased estimate of the execution cost which is directly related with the bid-ask spread. The Thompson and Waller measure is defined as,

$$TW = \frac{1}{T}\sum_{t=1}^{T}\left|\Delta p_t^*\right| \tag{7}$$

where $\Delta p_t^*$ is the series of non-zero price changes.

From (1) and (2), it can be shown that *TW* is a rough estimator of *c*. Under the assumption that zero price changes (i.e. $\Delta p_t = 0$) are a proxy for zero trade direction changes (i.e. $\Delta q_t = 0$), the expected value of $|\Delta p_t|$ is (see appendix):

$$E\left[\left|\Delta p_t^*\right|\right] = \frac{\sigma_\mu}{\sqrt{2\pi}} e^{-\frac{4c^2}{2\sigma_\mu^2}} + c\left[F(2c) - F(-2c)\right] \tag{8}$$

where *F(•)* is the normal cumulative distribution function with mean zero and variance $\sigma_\mu^2$.[3]

The *CFTC* measure is similar to the *TW* but eliminates any price change that follows another price change of the same sign. This modification is designed to reduce the chance of attributing variability due to new information to the bid-ask spread, and to eliminate serially correlated transaction types. The measure, described in Wang et al. (1990), is as follows,

$$CFTC = \frac{1}{T}\sum_{t=1}^{T}\left|\Delta p_t^{**}\right| \tag{9}$$

where $\Delta p_t^{**}$ is the series of non-zero price changes that are price reversals (i.e. subsequent price changes of different sign).

***Bayesian estimator***

Hasbrouck (2004) proposes Bayesian estimation to infer the effective bid-ask spread. Bayesian estimation is implemented using the Gibbs sampler which is a Markov chain Monte Carlo estimator. This technique is attractive because the estimation is based on parameters posteriors which incorporate all the information in the observed transaction prices, and the trade direction indicator is estimated conditional on observed transaction prices rather than assuming buys and sells are equally likely (as in *RM*) or derived from tick rules (as in *CDP*). Hasbrouck argues that an additional strength of the Bayesian estimate is that simulated posteriors are exact small sample distributions as they account for serial correlation and changes in information. However, as we show below, the correlation found in the estimation is inflated.

The Gibbs sampler generates a sequence of samples from the conditional probability distributions of random variables. The algorithm is motivated because it is applicable when the joint distribution $F(q,c,\sigma^2_u|p)$ is not known but the conditional distribution of each variable is known. As a Markov chain Monte Carlo method, the Gibbs sampler generates sample values from the distribution of each variable in turn, conditional on the current values of the other variables.

In the Bayesian approach the transaction cost, $c$, and the variance of the efficient price changes, $\sigma^2_u$, are the unknown parameters from the regression specification

$$\Delta p_t = c\Delta q_t + u_t \qquad u_t \sim N(0, \sigma^2_u) \qquad (10)$$

We use the Gibbs sampler to obtain sample values $(q^{(i)},c^{(i)},\sigma^2_u{}^{(i)}) \sim F(q,c,\sigma^2_u|p)$ based on known conditional distributions for a known set $p=\{p_1,p_2,...,p_T\}$. For the vector variable $\boldsymbol{\Theta}=(q,c,\sigma^2_u)$, a Markov chain is used to make $n$ random draws which converge in distribution to the joint distribution after a sufficiently large number of iterations. Notationally,

$$(q^{(0)},c^{(0)},\sigma^2_u{}^{(0)}),\ (q^{(1)},c^{(1)},\sigma^2_u{}^{(1)}),...,\ (q^{(n)},c^{(n)},\sigma^2_u{}^{(n)}) \qquad \boldsymbol{\Theta}^{(n)} \sim F^{(n)}(q,c,\sigma^2_u|p)$$

where $\boldsymbol{\Theta}^{(n)} \rightarrow^{D} \boldsymbol{\Theta}$ as $n \rightarrow \infty$. The liquidity cost $c$ is then computed as the first moment of the marginal distribution $f(c|p)$.

In Hasbrouck's approach, the conditional prior distribution for $c$ is imposed to be positive normal, so the posterior is $c|\,p \sim N^+(\mu_c^{post}, \Omega_c^{post})$, where, $N^+$ is the normal density restricted to $[0,+\infty)$, $\mu_c^{post}=Dd$, $\Omega_c^{post}=\sigma^2_u(X'X)^{-1}$, $D^{-1}=X'\sigma^2_u{}^{-1}X+(\Omega_c^{prior})^{-1}$, $d=X'\sigma^2_u{}^{-1}p+(\Omega_c^{prior})^{-1}\,\mu_c^{prior}$, $X=[\Delta q(t)]$, $\mu_c^{prior}= 0$, and $\Omega_c^{prior}=10^6$. The conditional posterior distribution for $\sigma^2_u$ is $\sigma^2_u \mid p \sim IG(\alpha^{post}, \beta^{post})$, where $\alpha^{post}= \alpha^{prior}+T/2$, and $\beta^{post}= \beta^{prior}+\Sigma u_t^2/2$, $\alpha^{prior} =\beta^{prior} =10^{-12}$.

The implementation of the algorithm follows a straightforward structure. Begin with an initial (arbitrary) guess of $(q,c,\sigma^2_u)^{(0)}$ and generate $n$ =1,000 draw sequences (we discard the first 20% considered as a burning time and keep the remaining 80% for estimation), where each draw incorporates the most recent information from previous draws and is conditional on the set of observed transaction prices $p$. Specifically,

1. Draw $c^{(1)}$ from $f(c|\sigma_u^{(0)}, q^{(0)}, p)$, $c|\,p \sim N^+(\mu_c^{post}, \Omega_c^{post})$
2. Draw $\sigma^2_u{}^{(1)}$ from $f(\sigma^2_u|m^{(0)}, p)$, $\sigma^2_u \mid p \sim IG(\alpha^{post}, \beta^{post})$
3. Draw $q^{(1)}$ from $f(q|c^{(1)}, \sigma_u^{(1)}, p)$, $q|c, \sigma_u, m, p \sim Bernoulli(p_{buy})$

   where $p_{buy}$ is the probability that $q$ = +1.[4]

Hasbrouck's spread estimates are considerable smaller than serial covariance estimates. Hasbrouck argues that this discrepancy is due to a small sample effect because the independence assumption in (10) is not imposed in the Bayesian model. Therefore, the term $E[u_t\, u_{t-1}]$ in (4) is not zero but usually negative. When this term is introduced in (4), solving for $c$ would yield lower Roll estimates that are similar to the Bayesian estimates. However, negative correlation in the error might not be due to a small sample effect as argued by Hasbrouck, but to the structure of the model. Specifically, the truncation of the distribution of the half spread $c$ might introduce correlation in the model. In the next section we analyze the consequences of truncating the distribution of $c$ and we derived an improved estimator.

### *Estimation of the half spread*

The estimation of the half spread $c$ deserves special attention as it is suspected that the adjustments done to meet the economic meaning of $c$ might actually bias its estimation. Specifically, to ensure that $c$ is positive, Hasbrouck limits its posterior distribution to positive values only. However, this truncation might introduce serial correlation between the errors $u_t$. To show this, we compute the serial correlation for non-truncated and truncated models. In view of the non-zero correlation found in the truncated case, we then suggest an alternative estimator of $c$ that uses the absolute values of $\Delta q_t$ and $\Delta p_t$. This is done for the simple case in which $T = 1$. Then we analyze the case for $T > 1$, for which the correlation appears to be non-zero even when no modification is done in the estimation of $c$ (i.e., no truncation is made). We then derive the conditions required for an estimator to yield zero serial correlation. These conditions appear to be very restricted. Finally, we impose a less restrictive correlation condition and we derive an estimator meeting this condition.

From the Roll model in (10),

$$c = \frac{\Delta p_t - u_t}{\Delta q_t} \qquad \Delta q_t \neq 0 \tag{11}$$

Then, it is straightforward that the probability density distribution for $c$ for any fixed $\Delta p_t$ and $\Delta q_t$ ($\Delta q_t \neq 0$) is[5],

$$c \left| \Delta p_t, \Delta q_t = \pm 2 \sim N\left( \frac{\Delta p_t}{\Delta q_t}, \frac{\sigma_u}{\left|\Delta q_t\right|} \right) \right. \tag{12}$$

In the Gibbs sampler the objective is to make draws from $c \mid \Delta p_t, \Delta q_t$. There are two cases for $\Delta q_t$, $\Delta q_t = -2$ and $\Delta q_t = 2$; and also $\Delta p_t$ could be $\Delta p_t > 0$ or $\Delta p_t < 0$. If $(\Delta p_t / \Delta q_t) > 0$, then $c$ is more likely to be positive and this will happen in approximately half of the loops of the Gibbs sampler. The final distribution of $c$ would look like Figure 1, from where it follows that the draw for $c$ could be positive or negative. However, because $c$ is associated with the cost of liquidity its value should be positive. Hasbrouck (2004) proposes a truncation, by drawing $c$ from a positive normal distribution. However, as shown below the truncation introduces correlation between $u_t$ and $u_{t+1}$. We first compute the correlation between consecutive errors when the draws for $c$ come from (12) and then we compute the correlation when the draws for $c$ come from the truncated distribution.

From the Roll model in (10), $u_t u_{t+1} = (\Delta p_t - c_t \Delta q_t)(\Delta p_{t+1} - c_{t+1} \Delta q_{t+1})$, where $c_t$ and $c_{t+1}$ are independent of each other. Distributing and taking expectations yields,

$$E[u_t u_{t+1} \mid \Delta p_t, \Delta p_{t+1}, \Delta q_t = \pm 2, \Delta q_{t+1} = \pm 2] =$$
$$= \Delta p_t \Delta p_{t+1} - E[c_t]\, \Delta q_t \Delta p_{t+1} - E[c_{t+1}]\, \Delta q_{t+1} \Delta p_t + E[c_t c_{t+1}]\, \Delta q_t \Delta q_{t+1} \tag{13}$$

Using (12) and substituting $E[c_t] = \Delta p_t / \Delta q_t$, $E[c_{t+1}] = \Delta p_{t+1} / \Delta q_{t+1}$ and $E[c_t c_{t+1}] = E[c_t] E[c_{t+1}] = (\Delta p_t / \Delta q_t)\, (\Delta p_{t+1} / \Delta q_{t+1})$ makes all the terms in (13) to cancel out and the resulting correlation between $u_t$ and $u_{t+1}$ to be zero.

When the distribution of $c$ is truncated and $c_t$ and $c_{t+1}$ are estimated independently, the posterior $E[u_t u_{t+1} | \Delta p_t, \Delta p_{t+1}, \Delta q_t = \pm 2, \Delta q_{t+1} = \pm 2]$ is no longer zero. To see this, take the expectation of $c_t$,

$$E[c_t] = \int_0^\infty \frac{c}{\sqrt{2\pi}\sigma} e^{-\frac{(c-\mu)^2}{2\sigma^2}} dc = \int_0^\mu \frac{c}{\sqrt{2\pi}\sigma} e^{-\frac{(c-\mu)^2}{2\sigma^2}} dc + \int_\mu^\infty \frac{c}{\sqrt{2\pi}\sigma} e^{-\frac{(c-\mu)^2}{2\sigma^2}} dc$$

letting $x = (c - \mu)^2$,

$$E[c_t] = -\int_0^{\mu^2} \frac{1}{2} \frac{e^{-\frac{x}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx + \int_0^\infty \frac{1}{2} \frac{e^{\frac{x}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dx + \int_0^\infty \mu \frac{e^{-\frac{(c-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dc$$

letting $w = c - \mu$ and solving,

$$E[c_t] = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left(1 - F(-\mu)\right) \qquad (14)$$

Using the truncated version of $E[c_t]$, (14), to solve (13) shows that $E[u_t u_{t+1} | \Delta p_t, \Delta p_{t+1}, \Delta q_t = \pm 2, \Delta q_{t+1} = \pm 2] \neq 0$. In Hasbrouck's model, the restriction $c| p \sim N^+(\mu, \Omega)$, comes from the prior c $\sim N^+(\mu^{prior}, \Omega^{prior})$, but this might conflict with the standard regression assumptions.

Another way to draw positive values for $c$ is to change the sign in $\Delta p_t/\Delta q_t$ by taking absolute values. The correlation in this case is,

$$E[u_t u_{t+1} | |\Delta p_t|, |\Delta p_{t+1}|, |\Delta q_t|, |\Delta q_{t+1}|] = - |\Delta p_t| |\Delta p_{t+1}| + E[c_t c_{t+1}] |\Delta q_t| |\Delta q_{t+1}|] = 0$$

By using this approach there is still a probability that $c < 0$ equal to the shaded area in Figure 1, however this area is negligible.

Next, we expand these results for the entire sample set. For $\Delta p = (\Delta p_1, \ldots, \Delta p_T)'$, $\Delta q = (\Delta q_1, \ldots, \Delta q_T)'$ and $u = (u_1, \ldots, u_T)'$, solving for $c$ in the Roll model yields,

$$c = \left(\Delta q' \Delta q\right)^{-1} \left(\Delta q' \Delta p - \Delta q' u\right) = \frac{\Delta q' \Delta p}{\Delta q' \Delta q} - \frac{\Delta q' u}{\Delta q' \Delta q} \qquad (15)$$

From (15) it follows that $c | \Delta p, \Delta q, \sigma_u \sim N\left(\frac{\Delta q' \Delta p}{\Delta q' \Delta q}, \frac{\sigma_u}{\Delta q' \Delta q}\right)$ (16)

which is the OLS estimator. It should be noted that $\Delta q_t = 0$ does not contribute neither to the mean $\mu = (\Delta q_1 \Delta p_1 + \ldots + \Delta q_T \Delta p_T)/(\Delta q_1^2 + \ldots + \Delta q_T^2)$ nor to the variance $\sigma^2 = \sigma_u^2/(\Delta q_1^2 + \ldots + \Delta q_T^2)$. We can generalize our previous result of using absolute values for the entire sample ($T > 1$), $\mu = (|\Delta q_1| |\Delta p_1| + \ldots + |\Delta q_T| |\Delta p_T|)/(\Delta q_1^2 + \ldots + \Delta q_T^2)$ and

$$\sigma = \sigma_u / \sqrt{\Delta q_1^2 + \ldots + \Delta q_T^2}$$

The correlation between subsequent errors for the entire sample differs from the simple case above,

$$E[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u] = E[(\Delta p_t - c\Delta q_t)(\Delta p_{t+1} - c\Delta q_{t+1})]$$
$$= \Delta p_t \Delta p_{t+1} - E[c] \Delta q_t \Delta p_{t+1} - E[c] \Delta q_{t+1} \Delta p_t + E[c^2] \Delta q_t \Delta q_{t+1}$$
$$= \Delta p_t \Delta p_{t+1} + \mu (\Delta q_t \Delta p_{t+1} + \Delta q_{t+1} \Delta p_t) + (\sigma^2 + \mu^2) \Delta q_t \Delta q_{t+1} \neq 0 \qquad (17)$$

The above result shows that, for the general case $T > 1$, the correlation does not vanish. Notice that for $T \to \infty$, $\mu \to 0$ because $\Delta q' \Delta p$ is expected to converge to zero[6] and $\Delta q' \Delta q$ to $+\infty$, and $\sigma^2 \to 0$ because $\sigma_u^2$ is fixed and $\Delta q' \Delta q \to +\infty$. Then,

$$E\left[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u\right] \xrightarrow{T \to \infty} \Delta p_t \Delta p_{t+1}$$

Furthermore, if $\Delta q_t \neq 0$ and $\Delta q_{t+1} \neq 0$ then for a finite sample,

$$E\left[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u\right] = \Delta p_t \Delta p_{t+1} + \left[\sigma^2 + \mu^2 - \mu\left(\frac{\Delta p_t}{\Delta q_t} + \frac{\Delta p_{t+1}}{\Delta q_{t+1}}\right)\right] \Delta q_t \Delta q_{t+1} \qquad (18)$$

In the above expression (18), as $T$ increases, $\sigma^2$ and $\mu^2$ become negligible with respect to $\mu$ because they are of quadratic order. Therefore, $\sigma^2$ and $\mu^2$ may be dropped for large values of $T$ and the correlation is,

$$E[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u] \cong \Delta p_t \Delta p_{t+1} - \mu (\Delta q_t \Delta p_{t+1} + \Delta q_{t+1} \Delta p_t)$$

For $\Delta q_t = \pm 2$ and $\Delta q_{t+1} = \pm 2$, $E[u_t u_{t+1}]$ will be distributed symmetrically on both sides of $\Delta p_t \Delta p_{t+1}$. If $p_{t-1} \approx p_{t+1}$, the time series behaves as a "sawlike" manner (i.e., negatively correlated as in Figure 2) and it is easy to see $\Delta p_t \Delta p_{t+1} = (p_t - p_{t-1})(p_{t+1} - p_t) \cong -(p_{t+1} - p_t)^2 = -\Delta p_{t+1}^2 < 0$.

We have shown above that for a sample size $T > 1$, that negative correlation will always exist Next, we try to answer the question, can the estimation of the half spread $c$ be improved? To answer this question, we analyze the conditions needed to have $E[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u] = 0$. Recall the posterior error correlation expression from (17) which can be written as,

$$E\left[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u\right] = \Delta q_t \Delta q_{t+1} \left[\frac{\Delta p_t}{\Delta q_t}\frac{\Delta p_{t+1}}{\Delta q_{t+1}} - \mu\left(\frac{\Delta p_t}{\Delta q_t} + \frac{\Delta p_{t+1}}{\Delta q_{t+1}}\right) + \left(\sigma^2 + \mu^2\right)\right]$$

for $\Delta q_t, \Delta q_{t+1} \neq 0$

Let $\mu_t = \Delta p_t / \Delta q_t$ and $\mu_{t+1} = \Delta p_{t+1} / \Delta q_{t+1}$, which are the means for $c_t$ and $c_{t+1}$ respectively. Then, the condition for $E[u_t u_{t+1} | \Delta p, \Delta q, \sigma_u] = 0$ is,

$$(\mu_t - \mu)(\mu_{t+1} - \mu) + \sigma^2 = 0$$

If $T \rightarrow \infty$, then $\sigma^2 \rightarrow 0$ and the condition for zero correlation is $\mu = \mu_t$ or $\mu = \mu_{t+1}$. For a sample containing $t = 1, 2, \ldots, T$ prices this appears to be a very restrictive condition as it must be fulfilled for all pairs $\mu_t\ \mu_{t+1}$ simultaneously. That is, $\mu$ should be simultaneously equal to $\mu_1$ *or* $\mu_2$, $\mu_2$ *or* $\mu_3, \ldots, \mu_{T-1}$ *or* $\mu_T$. Imposing no correlation between consecutive posterior errors simultaneously leads to imposing *T-1* conditions simultaneously, which in practice is not achievable. A less restrictive condition is,

$$E[u_1\, u_2\, u_3 \ldots u_T | \Delta p, \Delta q, \sigma_u] = 0 \qquad (19)$$

which is that the joint correlation for all errors is zero and leads to imposing only one condition instead of $T - 1$ simultaneous conditions. The condition for (19) is derived as follows,

$u_1\, u_2\, u_3 \ldots u_T = (\Delta p_1 - c\Delta q_1) \ldots (\Delta p_T - c\Delta q_T) = \Delta q_1 \ldots \Delta q_T\, (\mu_1 - c) \ldots (\mu_T - c)$
for $\Delta q_t, \Delta q_{t+1} \neq 0$

$$E[u_1\, u_2\, u_3 \ldots u_T | \Delta p, \Delta q, \sigma_u] = \pm 2^T E[(\Delta\mu_1 \ldots \Delta\mu_T) - (c - \mu)(\Delta\mu_1 \ldots \Delta\mu_{T-1} + \ldots + \Delta\mu_2 \ldots \Delta\mu_T) + (c - \mu)^2 (\Delta\mu_1 \ldots \Delta\mu_{T-2} + \ldots + \Delta\mu_3 \ldots \Delta\mu_T) + \ldots + (-1)^{\mathrm{T}} (c - \mu)^T \qquad (20)$$

Distributing the $E[.]$ operator among the sums we arrive at an expression containing $E[c - \mu] = 0$, $E[(c - \mu)^2] = \sigma^2, \ldots, E[(c - \mu)^T]$. From (17) we know that $c$ is normally distributed, therefore,

$$E\left[(c-\mu)^n\right] = \begin{cases} 0 & \text{for } n \text{ odd} \\ \dfrac{(2n)!}{2^n n!}\sigma^{2n} & \text{for } n \text{ even} \end{cases} \qquad (21)$$

From (16) we also know that $\sigma^2 = \dfrac{\sigma_u^2}{\Delta q_1^2 + \ldots + \Delta q_T^2} = \dfrac{1}{T}\dfrac{\sigma_u^2}{4} = \dfrac{1}{T}\left(\dfrac{\sigma_u}{2}\right)^2$ (22)

For any fixed $n$, as $T \rightarrow \infty$, $\sigma^2$ in (22) goes to zero, making $E[(c - \mu)^n$ in (21) go to zero. However, when $n = T$ this result is different,

$$\lim_{T\to\infty} \frac{(2T)!}{2^T T!}\sigma^{2T} = \lim_{T\to\infty} \frac{(2T)!}{2^T T!}\frac{1}{T^T}\left(\frac{\sigma_u}{2}\right)^{2T} = \lim_{T\to\infty}\left(1 - \frac{1}{2T}\right)\ldots\left(\frac{1}{2} + \frac{1}{2T}\right)\left(\frac{\sqrt{2}\sigma_u}{2}\right)^{2T}$$

In the above expression, $\lim_{T\to\infty}\left(1 - \dfrac{1}{2T}\right) = 1$, however $\lim_{T\to\infty}\left(\dfrac{\sqrt{2}\sigma_u}{2}\right)^{2T} = \begin{cases} 0 & \sigma_u < \sqrt{2} \\ \infty & \sigma_u > \sqrt{2} \end{cases}$

So, for $\sigma_u < \sqrt{2}$, (20) simplifies to,

$$E[\mu_1\, \mu_2\, \mu_3 \ldots \mu_T | \Delta p, \Delta q, \sigma_u] \cong \pm 2^T (\Delta\mu_1 \ldots \Delta\mu_T) = \pm 2^T (\mu_1 - \mu) \ldots (\mu_T - \mu)$$

If we now impose (19), zero correlation between errors, then ($\mu_1 - \mu$) … ($\mu_T - \mu$) $\cong$ 0 and $\mu$ = $\mu_1, \mu_2, ..., \mu_T$. That is, by choosing any $\mu_T$ the joint expectation of $\mu_1$ ... $\mu_T$ will vanish. However, this will introduce a bias in $\mu$ because the original $\mu$ was,

$$\mu = \frac{\Delta q_1 \Delta p_1 + ... + \Delta q_T \Delta p_T}{\Delta q_1^2 + ... + \Delta q_T^2} = \frac{\Delta q_1^2}{\Delta q_1^2 + ... + \Delta q_T^2}\mu_1 + ... + \frac{\Delta q_T^2}{\Delta q_1^2 + ... + \Delta q_T^2}\mu_T$$

$$\mu = \frac{\mu_1 + ... + \mu_T}{T} \tag{23}$$

The mean value in (23) minimizes the distances $d_t = |\mu_T - \mu|$ because

$$\frac{d}{d\mu}\left[(\mu_1 - \mu)^2 + ... + (\mu_T - \mu)^2\right] = 0 \Rightarrow \mu = \frac{\mu_1 + ... + \mu_T}{T}$$

So, choosing as $\mu$ the value $\mu_t$ that is closest to $(\mu_1 + ... + \mu_T)/T$ might provide a good estimator.

We can summarize the results obtained in this section as follows: 1) Truncating the distribution of $c$ introduces correlation between consecutive errors as shown by (14); 2) In the simple case $T$ = 1 the correlation between two consecutive errors introduced by the truncation is corrected when we take absolute values, $|\Delta q_t|$ and $|\Delta p_t|$; 3) When we expand the results for the entire sample set the correlation between two consecutive errors never vanishes as shown in (17). Moreover, the correlation will be approximately $\Delta p_t \Delta p_{t+1}$ and likely to be negative; 4) Imposing no correlation between two consecutive posterior errors leads to imposing $T$-$1$ conditions simultaneously, which in practice is not achievable, 5) A less restrictive condition is to impose no overall correlation between posterior errors which leads to the condition $\mu = \mu_1$ or $\mu_2$ or … $\mu_T$; and 6) The value $\mu_t$ closest to the usual estimator $(\mu_1 + ... + \mu_T)/T$ might provide a good estimator. However, no matter how close $\mu_t$ and $(\mu_1 + ... + \mu_T)/T$ are, convergence will only be possible if $\sigma_u < \sqrt{2}$ .

## Data

### *Simulated data*

We first use simulated price data to evaluate the different measures of liquidity costs. We simulate transaction prices using the basic Roll model as described in (1) and (2) which is the underlying model in all the bid-ask spread measures described above. We perform $k$ = 1,000 simulations, and generate a distribution of the half bid-ask spread estimates yielded by each spread estimator. For each simulation we generate $T$ = 500 prices which is within the range of the number of trades per day observed in the live cattle and lean hogs markets (Table 2). Simulated data follows the standard Roll model with the assumptions described above. This means that price change series with positive covariance were eliminated and estimators are assessed by their ability to reflect the Roll model under its most favorably conditions.

We simulate prices for different market conditions to study spread estimator response and the source of bias they might have. The parameters of the model are selected based on price behavior of the markets under study. In the model, the parameters are *c*, $q_t$ and $\sigma^2_u$ and are selected among the range of values reported in previous studies. For *c*, Locke and Venkatesh's (1997) found 1.5 and 3.6 for LC and LH respectively. Ferguson and Mann (2001) report values of 0.8 and 2 for LC and LH. We select $c = 1$ for the simulations as it seems to be a reasonable half spread value for LC and LH markets. For $q_t$, we use the default distribution *Bernoulli(1/2)*. We also use a higher probability of 0.7 as reported by Choi, Salandro and Shastri (1988) for options traded in the Chicago Board Options Exchange (CBOE) and a lower value of 0.3 (Chu, Ding, and Pyun 1996 report values between 0.4 and 0.5 in foreign exchange futures prices).[7] The variance $\sigma_u^2$ represents the level of noise in the market. In the simulation we represent three scenarios, with corresponding variances of 0.5, 1.0 and 1.5. In general, these values of $\sigma^2_u$ are high when compared with the level of *c*. We choose these values following results reported by Hasbrouck (2004) for pork bellies and preliminary analysis of our own data.

***Market data***

The simulated data is used to assess spread estimators' performance under controlled conditions in an ideal market. However, in practice, the market might behave differently with the introduction of new information. Real transaction prices differ from simulated prices in that the former might have, for example, short-term price trends, or substantial changes due to new information arrival, or large jumps due to infrequent trading. Therefore, we compare spread estimates using real data.

Liquidity costs in the real market are estimated for lean hogs and live cattle futures contracts. We choose these commodities as they are among the most traded agricultural commodities in the CME and therefore are of interest for many market participants. We use the *volume by tick* database from the CME, which provides prices of all trades (including zero price changes) executed during the day in the open auction with their corresponding time stamps.

Preliminary analysis of the data shows that the spread estimators described in the previous section are sensitive to the time interval used. Roll (1984) finds differences between spreads estimated from daily and weekly data in the stock market. He suggests that this difference could be caused by market inefficiencies or non-stationary data. Non-stationarity may be due to short-term fluctuations (in expected returns) which dampen out over longer periods. The spread itself might be non-stationary due to the reaction of dealers to stochastic information arrival. Stationarity of the data is hard to assess because transaction prices are observed at unequally spaced intervals. Therefore, following common practice in the literature (e.g. Bryant and Haigh 2004, Locke and Venkatesh 1997) to avoid estimation problems due to non-stationary price behavior, we estimate liquidity costs on a daily basis.

For each commodity we selected three contracts with differing trading activity. Table 1 summarizes the trading month and contract specifications, and Table 2 shows summary descriptive statistics for each contract.

## Results

Table 3 shows the response of each spread estimator to simulated data for different scenarios of increasing levels of noise, $\sigma_u$, and probabilities of trade directions $q$, $p_{buy}$. Table 4 shows the correlation coefficients for all spread estimators and corresponding scenarios of noise in Table 3. For clarity of exposition we only included the correlation coefficients for the base case when the $p_{buy} = 0.5$. The *RM* estimator is included in the first line of each scenario and as expected is able to replicate itself when prices are generated in line with all its five assumptions. When the fifth assumption is not met, that is, when the probability of each transaction being a purchase is different from than that of being a sale (in the table, $p_{buy} \neq 0.5$), the expected value of the *RM* estimator for 1,000 simulations is slightly downward biased, however the true value of $c = 1$ falls inside the two SD interval for the range of noise in all simulations (from $\sigma_u = 0.5$ to $\sigma_u = 1.5$). Also as expected, the *RM* estimator shows increasing precision as the level of noise declines (i.e., lower SD and MSE as $\sigma_u$ gets lower).

The top portion of the tables corresponds to a market with a low level of noise. Under this scenario, the Gibbs sampler estimators (*HAS, ABS,* and *AVG*) seem to perform better that other measures. The level of precision of these estimators when $p_{buy} = 0.5$ is higher than the *RM* measure, a reflection of the limited amount of noise and the iterative nature of the Gibbs sampler estimator which uses 1,000 iterations to estimate $c$ for each series of price changes. When $p_{buy} \neq 0.5$ the Gibbs sampler estimates behave much like the *RM* estimate, with a lower expected mean but containing the true value in their two SD interval. From all these three measures, *ABS* and *AVG* behave similarly, however *HAS* shows a larger downward bias together with a higher SD and MSE. Surprisingly, the correlation between *AVG* and all other spread estimators (including ABS) is practically zero.

For the other estimators, the half spread $c = 1$ does not fall inside their two SD interval. *CDP* is downward biased and the mean absolute price change estimators are upward biased. The *CFTC* has the highest bias and SD, however both mean absolute price change estimates (*CFTC* and *TW*) are closer to the generated half spread of one when $p_{buy} \neq 0.5$. The correlation coefficients are higher between these estimators and *RM* than those between *HAS* or *ABS* and *RM*. However, because this is the scenario with the lower level of noise in the market, correlation coefficients may not be very informative as it is likely that most estimators perform well under these conditions.

The second part of Table 3 represents a market in which the level of noise $\sigma_u$ equals the parameter to be estimated, $c$. As expected, the performance of the spread estimators decreases. All estimators but *CFTC* and *TW* show even a lower performance for data generated with a $p_{buy} \neq 0.5$. As in the top part of the table, *CFTC* and *TW* estimates are less upward biased and have a lower MSE for both cases in which the probability of incoming trades is different for buys and sells. However, all spread estimators other than *RM* do not contain the generated $c = 1$. It is important to note that *ABS* and *AVG* yield similar estimates which are the ones with the highest performance relative to *RM* and their correlation remains very low. The *CDP* estimator also has a low MSE and is highly correlated with *RM*, as it is expected as both are serial covariance estimators. However, it doesn't seem to be capturing the different buy/sell probabilities as we expected based on its design.

When the level of noise increases to a higher level than the parameter *c*, as shown in the third part of Table 3, all spread estimators but *ABS* and *AVG* get more inaccurate. The relative performance follows the same pattern as just described, *ABS* and *AVG* are the most accurate estimates (given by their estimate of *c* of 0.97 and low SD and MSE), followed by *CDP* and *HAS* which have lower MSE than *TW* and *CFTC* when $p_{buy} = 0.5$. The *RM*, *CDP* and *HAS* estimators have correlation coefficients higher than 0.90 while *TW* and *CFTC* are more highly correlated. In all cases *CFTC* shows the highest MSE. The high correlation between *ABS* and *TW* might be due the fact that both use absolute values in the estimation. The picture changes slightly when the prices are generated with a different $p_{buy}$. The mean absolute price change estimates and their precision slightly improve and all other estimates become more downward biased.

The results for all three scenarios of $\sigma_u$ show that when the data is generated with a $p_{buy}$ different from 0.5, the Gibbs sampler and serial covariance estimates of *c* become more downward biased and less precise. For the Gibbs sampler estimates these results were not expected as $P(q_t = 1) = 0.5$ is not a necessary assumption as it is for the *RM* estimator. Furthermore, the Gibbs sampler yields estimates of the latent variable $q_t$ instead of using tick rules. Table 5 shows the estimated $p_{buy}$ ( $\hat{p}_{buy}$ ) for the above scenarios. When $\sigma_u$ is low, the estimated $p_{buy}$ is closer to the generated $p_{buy}$ (for example, $\hat{p}_{buy} = 0.35$ for $p_{buy} = 0.30$) and these results are similar for all Gibbs sampler estimators and values of $p_{buy}$. However, when $\sigma_u$ increases, the estimated $p_{buy}$ converges to 0.50. These results can be explained using the expression for $p_{buy}$ in footnote 4. Both the numerator and denominator basically contain *e* raised to a ratio between *c* and $\sigma_u$. Keeping *c* constant at one, as $\sigma_u$ increases the ratio goes to zero and each exponential goes to one which gives the result of 0.5. Therefore, in markets with a high level of noise accurate estimates of $q_t$ are hard to obtain and they are expect to be close to 0.5. For the serial covariance estimators, *CDP* is basically the *RM* estimator with the assumption of $P(q_t = 1) = 0.5$ relaxed. The correction is made through $\delta$ which is the probability that a transaction at time *t* is of the same type than the transaction at *t-1*, and $\alpha$ which takes into account one more period. Table 5 shows the estimated $\delta$ for each scenario. The estimated $\delta$ cannot be directly compared with $p_{buy}$. Intuitively, $\delta$ can be thought as the probability that a $q_t = 1$ follows a $q_{t-1} = 1$ or that a $q_t = -1$ follows a $q_{t-1} = -1$. That is, $P(q_t = 1)$ $P(q_{t-1} = 1) + P(q_t = -1)\ P(q_{t-1} = -1) = P(q_t = 1)^2 + P(q_t = -1)^2 = p_{buy}{}^2 + (1 - p_{buy})^2$. Therefore, a generated $p_{buy}$ of 0.3 or 0.7 could be associated with a $\delta$ of 0.58. When the level of noise in the market is low (i.e., $\sigma_u = 0.5$), the estimated $\delta$ follows this behavior closely (i.e., $\hat{\delta} = 0.54$). However, *q* is a latent variable which is not known and *CDP* uses the tick rule based on price changes instead of $q_t$ to determine the direction of the trade. Therefore, when the level of noise is high, the estimated $\delta$ might not yield these exact values based on $q_t$. Furthermore, when the level of noise in the market is high and prices fluctuate more, it is likely that $q_t$ goes from +1 to -1 more frequently, and the probability of two consecutive $q_t$ being the same is low. Therefore, it is expected that for large $\sigma_u$, $\delta$ reaches a minimum value, which in terms of $q_t$ would be 0.5[8], but in terms of $p_t$ could be lower. In fact, for the generated data in Table 5

$\hat{\delta}$ converges to approximately 0.3 regardless the value of $p_{buy}$ for which the data was generated.

Next, we analyze a common issue found in prices which is the presence of correlation, and its effects on the estimates of the half spread. Table 6 shows the estimates of $c$ for data generated with different levels of correlation ($\rho$) in the error term $u_t$ of the Roll model and for the same three same scenarios of noise as before. First we compute the correlation coefficient of $u_t$ in the Roll model of the Gibbs sampler for the simple case in which data are generated with *iid* $u_t$. The coefficient of correlation for *HAS*, *ABS* and *AVG* are negative and oscillate between -0.1 and -0.2 depending on the level of noise in the data. These results are in line with our theoretical finding that for a sample size of $T > 1$ there will always exist negative correlation as shown by (17). Then we generated data using two different levels of correlation, $\rho$ = -0.2 and $\rho$ = -0.4 and compare these results with the *iid* case. For comparison purposes and ease of exposition, in Table 6 we repeat the estimates of $c$ from Table 3 which correspond to the base case $\rho$ = 0.

The top portion of Table 6 shows that when $\sigma_u$ is low, the generated correlation does not have a great impact on spread estimates and the estimation of $\rho$ is fair relative to the actual $\rho$. However, as $\sigma_u$ increases the serial covariance estimators become upward biased for increasing levels of $\rho$. This behavior can be explained using (4). Serial covariance estimates assume that the first term, $E[u_t\, u_{t-1}]$, is zero when in this case should be negative. Therefore, in the presence of negatively correlated data which is likely to occur due to the bid-ask bounce, serial covariance estimates will be inflated. When $\rho$ = 0 this is not a problem and both *RM* and Gibbs sampler estimators yield similar estimates. In all cases the estimated $\rho$ roughly approximates the actual $\rho$.

The Gibbs sampler estimators show a different response to correlated data. For the lowest level of $\sigma_u$ there are no differences between estimators and for different levels of $\rho$. As $\sigma_u$ increases, the level of correlation has some impact on c estimates. The impact is greater for the *HAS* estimator and it is more pronounced for the higher level of $\sigma_u$ as $E[\hat{c}]$ increases from 0.68 to 0.88 for $\rho$ = 0 and $\rho$ = -0.4 respectively. For the same level of noise, *ABS* and *AVG* only change from 0.97 to 1.01. The mean absolute price change estimators doesn't seem to be affected by the level of correlation introduced in the data. Particularly, *TW* remains unchanged for the three levels of $\rho$, 0, -0.2, and -0.4. The CFTC slightly increases for higher $\rho$ when the noise increases. However these estimators are upward biased for levels of noise and $\rho$. These results indicate that for a market in which the prices show a fair level of negative correlation *ABS* and *AVG* are the best spread estimators.

The analysis with the market data was developed to provide insights into the performance of the estimators in relevant market situations. Initially, we investigate the behavior of the estimators prior to and in the maturity month which among other things will provide insights into the presence of a maturity effect (Do the estimators provide different information regarding liquidity cost behavior as maturity approaches?). Then, assuming that traders do not hold contracts into the maturity month which is common practice, we examine the differences in liquidity costs from trading in the nearby contract (which is usually the most actively traded contract) and two subsequent contracts.

To examine the maturity effect, Figure 3 provides estimates of daily liquidity costs starting 50 days prior to maturity for the April cattle and October hog contracts (other contracts yield similar pictures). For ease of exposition, we included one serial covariance estimator (*RM*), one mean absolute price change estimator (*TW*), the Bayesian estimator (*HAS*), and its proposed modification (*ABS*). In the figures, spread estimators behavior is consistent with the findings using simulated data in the presence of $\rho$<0. The average estimated $\rho$ are -0.25 and -0.29 for live cattle and lean hogs respectively. The average estimated $\sigma_u$ are 6.30 and 6.09 for live cattle and lean hogs respectively which is a little more than twice the estimated $c$. Based on these results, we expect that the spread estimators behave as in the last portion of Table 6. The *HAS* estimator generates measures of liquidity costs that are almost always considerably smaller relative to the other estimators. The *ABS* estimator is always higher than *HAS* and lower than the rest of the estimators, and just like the simulated case, it might be the least biased with respect to actual spreads. The *RM* estimates are higher than *ABS* which might be due to the failure to account for the negative correlation. The *TW* estimator is the highest and it is occasionally exceeded by *RM*. With regards to the maturity effect, no clear evidence emerges to support a sharp increase in liquidity costs in the expiration month for any of the estimators or contracts.

To examine whether there are differences in liquidity costs from trading in the nearby contract, we compare the liquidity costs for nearby and subsequent contracts generated on the same day in the contract prior to maturity. For example, Figure 4 a) shows the liquidity cost estimates for the live cattle February, April, and June contracts generated daily in January, one month prior to nearby maturity. Similarly, Figure 4 b) shows the liquidity costs estimates for the lean hog August, October, and December contracts generated daily one month prior to nearby maturity in July. Liquidity cost estimates are provided for the *ABS* estimator as it has shown to be the best choice for the particular conditions of these markets.

Figure 4 a) suggests that there is no real difference in liquidity cost from trading in the nearby and more distant contracts for live cattle. This is somewhat surprising because traders often profess to trade in the nearby contract because of higher volume and low costs of liquidity. For lean hogs, Figure 4 b) suggests that trading in more distant contracts is associated with a higher liquidity cost.

## Conclusions

Estimating liquidity costs in agricultural futures markets is challenging because bids and asks occur in an open outcry and are not recorded. The problem becomes more severe as different measures that use transaction prices have been reported to be biased with respect to actual bid-ask spreads. Here we assess the performance of different conventional measures, a recently proposed Bayesian estimator, and our suggested modified estimator, using simulated and market data.

Different spread estimators may be more or less suitable to different market conditions. Serial covariance estimators (*RM* and *CDP*) cannot be computed when subsequent price changes have positive covariance. Absolute price change (*TW* and *CFTC*) and Bayesian (*HAS*) measures do not have this problem, however, the absolute price change measures might not distinguished true price change from bid-ask spread. The *HAS* estimator seems to be downward biased when the level of noise in the market is high and there are short term inefficiencies (i.e., negative correlation). To overcome this problem we propose a modification of the *HAS* estimator. Instead of truncating the distribution of *c* to make sure that the draw will yield positive values we suggest taking absolute values (*ABS* estimator).

For the simulated data, we perform the analysis for different market scenarios with increasing levels of noise. When the market is least noisy, we find that Gibbs sampler estimators (*HAS*, *ABS* and *AVG*) provide highly precise estimates of liquidity costs, even more precise than the *RM* measure estimated with simulated data entirely consistent with its structure. This finding is likely attributable to the iterative procedure used to develop these estimates. As the noise in the market increases (i.e., higher $\sigma_u$), the more biased the estimators become. An exception is the *ABS* estimator which seems to be the more appropriate under common conditions in the real market. Absolute price changes measures (*TW* and *CFTC*) appear to be upward biased, while the serial covariance (*CDP*) and Bayesian (*HAS*) measures are downward biased. The magnitude of the bias as reflected by the MSE is largest for the absolute price change measures in all scenarios.

However, when we use real data, market effects such as serial correlation and changes in information emerge that are captured by the *ABS* estimator and not by the rest of the estimators. Using market data we also estimated liquidity costs for nearby and distant contracts and for different days prior to expiration. We found no differential behavior of the estimators for different contracts in live cattle. For lean hogs more distant contracts appear to have higher liquidity costs. In both commodities we found no expiration effects for the contracts and periods analyzed.

## References

Anand, A. and A. Karagozoglu. “Relative Performance of Bid-Ask Spreads Estimators: Futures Markets Evidence.” *Journal of International Financial Markets, Institutions and Money* 16(2006): 231-245.

Brorsen, W., D. Buck, and S. Koontz. “Hedging Hard Red Winter Wheat: Kansas City versus Chicago.” *Journal of Futures Markets* 18(1998):449-466.

Bryant, H., and Michael Haigh. “Bid-Ask Spreads in Commodity Futures Markets.” *Applied Financial Economics* 14(2004):923-936.

Choi, J.Y, Dan Salandro and Kuldeep Shastri. “On the Estimation of Bid-Ask Spreads: Theory and Evidence.” *Journal of Financial and Quantitative Analysis* 23(2): 219-229
Chu, Q. C., Ding, D. K. and C. S. Pyun 1996. “Bid-ask and Spreads in the Foreign Exchange Market.” *Review of Quantitative Finance and Accounting* 6(1988):19-37.

Ferguson, M. and S. Mann. “Execution Costs and their Intraday Variation in Futures Markets.” Journal of Business 74(2001):125-160.

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

Lee, C. and M. Ready. “Inferring Trade Direction from Intraday Data.” *Journal of Finance* 46(1991):733-46.

Locke, P., and P.C. Venkatesh. “Futures Markets Transaction Costs.” *Journal of Futures Markets* 17(1997):229-245.

Ma, C., R. Peterson, and S. Sears. “Trading Noise, Adverse Selection, and Intraday Bid-Ask Spreads in Futures Markets.” *Journal of Futures Markets* 12(1992):519-538.

Pennings, J.M.E., and M. Meulenberg. “Hedging Efficiency: A Futures Exchange Management Approach.” *Journal of Futures Markets* 17(1997):599-615.

Phillips, S. and C. Smith. “Trading Costs for Listed Options: The implications for Market Efficiency.” *Journal of Financial Economics* 8(1980):179-201.

Roll, R. “A Simple Implicit Measure of the Effective Bid-ask Spread in an Efficient Market.” *Journal of Finance* 23(1984):1127-1139.

Smith, T., and R. Whaley. “Estimating the Effective Bid-ask Spread from Time and Sales Data.” *Journal of Futures Markets* 14(1994):437-456.

Thompson, S., and M. Waller. “The Execution Cost of Trading in Commodity Futures Markets.” *Food Research Institute Studies* 20(1987):141-163.

Thompson, S., and M. Waller. “Determinants of Liquidity Costs in Commodity Futures Markets.” *Review of Futures Markets* 7(1988):110-126.

Thompson, S., J. Eales, and D. Seibold. “Comparison of Liquidity Costs between Kansas City and Chicago Wheat Futures Contracts*.” Journal of Agricultural and Resource Economics* 18(1993):185-197.

Wang, H.K., E. Moriarty, R. Michalski, and J. Jordan. “Empirical Analysis of the Liquidity of the S&P 500 Index Futures Markets during the October 1987 Market Break.” *Advances in Futures and Options Research* 4(1990):191-218.

Wang, H.K., J. Yau, and T. Baptiste. “Trading Volume and Transaction Costs in Futures Markets.” *Journal of Futures Markets* 17(1997):757-78.

# Endnotes

[1] They compute the actual customer execution spread which is defined as the mean customer buy price less the mean customer sell price for 5-minute intervals. They use Computerized Trade Reconstruction (CTR) audit trail data provided by the Commodity Futures Trading Commission (CFTC). These data are not available to the public; it is only occasionally available for academic purposes. Ferguson and Mann study agricultural futures contracts traded in the Chicago Mercantile Exchange (CME).

[2] Bryant and Haigh use coffee and cocoa bid-ask quotes from the London International Financial Futures Exchange (LIFFE), and Anand and Karagozoglu use two financial futures bid-ask quotes from the Sydney Futures Exchange (SFE). Bid-ask quotes are not available from US exchanges.

[3] When $c >> \sigma_u$ (i.e., small random noise compared to $c$), $E\left[\left|\Delta p_t^*\right|\right]$ converges to $c$ because the exponential in the first term becomes negligible and $F(2c)-F(-2c)\cong 1$ (note that the integration limits $2c$ and $-2c$ are much higher than $2\sigma_\mu$ and $-2\ \sigma_u$ which is approximately 0.95). If the level of noise in the model is high, say $c \cong \sigma_u$, $E\left[\left|\Delta p_t^*\right|\right] \cong 1.004c$ because $F(2c)-F(-2c)\cong 0.95$ and the first term in (8) reduces to 0.054 $c$.

[4] The expression for $p_{buy}$ which is in Hasbrouck's (2004) appendix is: $\frac{e^{\frac{4cp_t}{\sigma_u^2}}}{e^{\frac{2c(m_{t-1}+m_{t+1})}{\sigma_u^2}}+e^{\frac{4cp_t}{\sigma_u^2}}}$

[5] The case $\Delta q_t = 0$ is not possible because that would make $u_t = \Delta p_t$, making $u_t$ fixed at $\Delta p_t$ which is not consistent with $u_t$ ~N(0, $\sigma_u$). It follows that $\Delta q_t = 0$ is non informative.

[6] Positive values of $\Delta q$ and $\Delta p$ are equally likely. Because $\mu \rightarrow 0$, it is necessary to make a truncation or take absolute values to estimate $c$.

[7] To our knowledge, there are no studies assessing the conditional probability of the transaction type in commodity futures markets. These two studies' estimates are upper and lower values of the default probability of 0.5 that is usually assumed. Therefore, they constitute a good reference for our purposes of studying spread estimators response to different market conditions.

[8] $\frac{\partial\delta^*}{\partial p_{buy}} = \frac{\partial}{\partial p_{buy}}\left[p_{buy}^2 + (1-p_{buy})^2\right] = 2p_{buy} - 2(1-p_{buy}) = 0 \Rightarrow p_{buy} = 0.5$, where δ* represents an intuitive approximation of $\delta$ based on $q_t$ (if they were known) rather than in the differences of $p_t$.

**Table 1: Trading month and contract specifications.**

| | **Live Cattle** | **Lean Hogs** |
|---|---|---|
| Trading month | Jan 05 | Jul 05 |
| Tick | 0.025 | 0.025 |
| Size of contract | 40,000 lb | 40,000 lb |

**Table 2: Summary descriptive statistics**

| **Commodity** | | **Live cattle** | | | **Lean Hogs** | |
|---|---|---|---|---|---|---|
| Expiration month | Feb | Apr | Jun | Aug | Oct | Dec |
| Avg. price (cents/lb.) | 89.78 | 87.95 | 82.02 | 67.25 | 58.17 | 55.64 |
| Standard deviation | 0.28 | 0.24 | 0.21 | 0.26 | 0.26 | 0.22 |
| Min price (cents/lb.) | 87.40 | 85.65 | 83.78 | 64.85 | 56.20 | 54.50 |
| Max price (cents/lb.) | 93.00 | 90.30 | 83.78 | 69.10 | 59.80 | 56.65 |
| Avg. daily volume | 11597 | 8798 | 2117 | 7725 | 5971 | 1007 |
| Avg. daily trades | 634.2 | 550.8 | 201 | 611 | 335 | 93.8 |
| Avg. time b/w trades (sec.) | 25.31 | 31.35 | 80.27 | 26.26 | 36.67 | 113.65 |

**Table 3: Estimates of the half bid-ask spread using simulated data**

a) Parameters of the simulated data: $c$ = 1 and $\sigma_u$ = 0.5

| | **$p_{buy}$ = 0.3** | | | **$p_{buy}$ = 0.5** | | | **$p_{buy}$ = 0.7** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** |
| *RM* | 0.92 | 0.05 | 0.01 | 1.00 | 0.05 | 0.00 | 0.92 | 0.05 | 0.01 |
| *CDP* | 0.88 | 0.04 | 0.02 | 0.89 | 0.03 | 0.01 | 0.88 | 0.04 | 0.02 |
| *TW* | 1.07 | 0.04 | 0.01 | 1.20 | 0.04 | 0.04 | 1.07 | 0.04 | 0.01 |
| *CFTC* | 1.23 | 0.06 | 0.06 | 1.36 | 0.05 | 0.13 | 1.23 | 0.06 | 0.06 |
| *HAS* | 0.85 | 0.14 | 0.04 | 0.98 | 0.02 | 0.00 | 0.85 | 0.14 | 0.04 |
| *ABS* | 0.91 | 0.09 | 0.02 | 0.98 | 0.02 | 0.00 | 0.91 | 0.09 | 0.02 |
| *AVG* | 0.90 | 0.09 | 0.02 | 0.98 | 0.02 | 0.00 | 0.91 | 0.09 | 0.02 |

b) Parameters of the simulated data: $c$ = 1 and $\boldsymbol{\sigma_u}$ **= 1**

| | **$p_{buy}$ = 0.3** | | | **$p_{buy}$ = 0.5** | | | **$p_{buy}$ = 0.7** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** |
| *RM* | 0.92 | 0.07 | 0.01 | 1.00 | 0.07 | 0.01 | 0.92 | 0.07 | 0.01 |
| *CDP* | 0.75 | 0.06 | 0.06 | 0.81 | 0.06 | 0.04 | 0.76 | 0.06 | 0.06 |
| *TW* | 1.31 | 0.04 | 0.10 | 1.41 | 0.04 | 0.17 | 1.31 | 0.05 | 0.10 |
| *CFTC* | 1.44 | 0.06 | 0.19 | 1.54 | 0.06 | 0.29 | 1.44 | 0.07 | 0.20 |
| *HAS* | 0.62 | 0.05 | 0.14 | 0.72 | 0.05 | 0.08 | 0.62 | 0.05 | 0.14 |
| *ABS* | 0.79 | 0.03 | 0.05 | 0.86 | 0.03 | 0.02 | 0.79 | 0.03 | 0.05 |
| *AVG* | 0.79 | 0.03 | 0.05 | 0.86 | 0.03 | 0.02 | 0.79 | 0.03 | 0.05 |

c) Parameters of the simulated data: $c$ = 1 and $\boldsymbol{\sigma_u}$ **= 1.5**

| | **$p_{buy}$ = 0.3** | | | **$p_{buy}$ = 0.5** | | | **$p_{buy}$ = 0.7** | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** | $E[\hat{c}]$ | **SD** | **MSE** |
| *RM* | 0.91 | 0.10 | 0.02 | 1.00 | 0.10 | 0.01 | 0.92 | 0.10 | 0.02 |
| *CDP* | 0.72 | 0.08 | 0.08 | 0.79 | 0.08 | 0.05 | 0.72 | 0.08 | 0.08 |
| *TW* | 1.59 | 0.05 | 0.35 | 1.66 | 0.05 | 0.44 | 1.59 | 0.05 | 0.35 |
| *CFTC* | 1.70 | 0.07 | 0.49 | 1.78 | 0.07 | 0.61 | 1.69 | 0.08 | 0.49 |
| *HAS* | 0.60 | 0.08 | 0.17 | 0.68 | 0.07 | 0.11 | 0.60 | 0.07 | 0.17 |
| *ABS* | 0.92 | 0.03 | 0.01 | 0.97 | 0.03 | 0.00 | 0.92 | 0.03 | 0.01 |
| *AVG* | 0.92 | 0.03 | 0.01 | 0.97 | 0.03 | 0.00 | 0.92 | 0.03 | 0.01 |

Notes: *RM* (Roll) and *CDP* (Chu, Ding, and Pyun) are serial covariance estimators, *TW* (Thompson and Waller) and *CFTC* (Commodity Futures Trading Commission) are mean absolute price change estimators, *HAS* (Hasbrouck), *ABS* and *AVG* are estimators using the Gibbs sampler. All estimations come from $k$ = 1,000 simulations using $T$ = 500 simulated prices, and $n$ = 1,000 iterations for the Gibbs sampler.

**Table 4: Correlation coefficients between *c* estimators using simulated data**

a) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{0.5}$

| | *RM* | *CDP* | *TW* | *CFTC* | *HAS* | *ABS* | *AVG* |
|---|---|---|---|---|---|---|---|
| *RM* | 1.00 | | | | | | |
| *CDP* | 0.75 | 1.00 | | | | | |
| *TW* | 0.83 | 0.54 | 1.00 | | | | |
| *CFTC* | 0.70 | 0.59 | 0.76 | 1.00 | | | |
| *HAS* | 0.32 | 0.41 | 0.45 | 0.35 | 1.00 | | |
| *ABS* | 0.30 | 0.40 | 0.44 | 0.32 | 0.82 | 1.00 | |
| *AVG* | 0.02 | -0.01 | 0.01 | -0.01 | -0.01 | -0.01 | 1.00 |

b) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{1}$

| | *RM* | *CDP* | *TW* | *CFTC* | *HAS* | *ABS* | *AVG* |
|---|---|---|---|---|---|---|---|
| *RM* | 1.00 | | | | | | |
| *CDP* | 0.90 | 1.00 | | | | | |
| *TW* | 0.67 | 0.52 | 1.00 | | | | |
| *CFTC* | 0.65 | 0.61 | 0.79 | 1.00 | | | |
| *HAS* | 0.84 | 0.76 | 0.77 | 0.66 | 1.00 | | |
| *ABS* | 0.70 | 0.63 | 0.92 | 0.75 | 0.89 | 1.00 | |
| *AVG* | -0.05 | -0.06 | -0.03 | 0.00 | -0.03 | -0.02 | 1.00 |

c) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{1.5}$

| | *RM* | *CDP* | *TW* | *CFTC* | *HAS* | *ABS* | *AVG* |
|---|---|---|---|---|---|---|---|
| *RM* | 1.00 | | | | | | |
| *CDP* | 0.95 | 1.00 | | | | | |
| *TW* | 0.50 | 0.41 | 1.00 | | | | |
| *CFTC* | 0.56 | 0.53 | 0.80 | 1.00 | | | |
| *HAS* | 0.93 | 0.88 | 0.55 | 0.57 | 1.00 | | |
| *ABS* | 0.63 | 0.59 | 0.94 | 0.79 | 0.70 | 1.00 | |
| *AVG* | 0.03 | 0.04 | -0.01 | 0.04 | 0.03 | -0.01 | 1.00 |

Notes: *RM* (Roll) and *CDP* (Chu, Ding, and Pyun) are serial covariance estimators, *TW* (Thompson and Waller) and *CFTC* (Commodity Futures Trading Commission) are mean absolute price change estimators, *HAS* (Hasbrouck), *ABS* and *AVG* are estimators using the Gibbs sampler. All estimations come from $k = 1{,}000$ simulations using $T = 500$ simulated prices with $p_{buy} = 0.5$, and $n = 1{,}000$ iterations for the Gibbs sampler.

**Table 5: Estimates of the $p_{buy}$ using simulated data**

a) Parameters of the simulated data: $c$ = 1 and $\boldsymbol{\sigma_u}$ **= 0.5**

| | $p_{buy}$ **= 0.3** | | $p_{buy}$ **= 0.5** | | $p_{buy}$ **= 0.7** | |
|---|---|---|---|---|---|---|
| | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** |
| *CDP* | $\hat{\delta}=0.54$ | 0.02 | $\hat{\delta}=0.47$ | 0.02 | $\hat{\delta}=0.54$ | 0.02 |
| *HAS* | 0.37 | 0.05 | 0.50 | 0.02 | 0.63 | 0.05 |
| *ABS* | 0.35 | 0.04 | 0.50 | 0.02 | 0.65 | 0.04 |
| *AVG* | 0.35 | 0.04 | 0.50 | 0.02 | 0.65 | 0.04 |

b) Parameters of the simulated data: $c$ = 1 and $\boldsymbol{\sigma_u}$ **= 1**

| | $p_{buy}$ **= 0.3** | | $p_{buy}$ **= 0.5** | | $p_{buy}$ **= 0.7** | |
|---|---|---|---|---|---|---|
| | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** |
| *CDP* | $\hat{\delta}=0.36$ | 0.03 | $\hat{\delta}=0.35$ | 0.03 | $\hat{\delta}=0.36$ | 0.03 |
| *HAS* | 0.49 | 0.00 | 0.50 | 0.00 | 0.51 | 0.00 |
| *ABS* | 0.48 | 0.01 | 0.50 | 0.01 | 0.52 | 0.01 |
| *AVG* | 0.48 | 0.01 | 0.50 | 0.01 | 0.52 | 0.01 |

c) Parameters of the simulated data: $c$ = 1 and $\boldsymbol{\sigma_u}$ **= 1.5**

| | $p_{buy}$ **= 0.3** | | $p_{buy}$ **= 0.5** | | $p_{buy}$ **= 0.7** | |
|---|---|---|---|---|---|---|
| | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** | $E[\hat{p}_{buy}]$ | ***SD*** |
| *CDP* | $\hat{\delta}=0.28$ | 0.04 | $\hat{\delta}=0.29$ | 0.03 | $\hat{\delta}=0.28$ | 0.04 |
| *HAS* | 0.50 | 0.00 | 0.50 | 0.00 | 0.50 | 0.00 |
| *ABS* | 0.49 | 0.00 | 0.50 | 0.00 | 0.51 | 0.00 |
| *AVG* | 0.49 | 0.00 | 0.50 | 0.00 | 0.51 | 0.00 |

Notes: $\delta$ is the probability that a transaction at time $t$ is of the same type than the transaction at $t$-$1$. $p_{buy}$ is the probability of $q_t = 1$. *CDP* (Chu, Ding, and Pyun) is a serial covariance estimator, *HAS* (Hasbrouck), *ABS* and *AVG* are Gibbs sampler estimators. All estimations come from $k$ = 1,000 simulations using $T$ = 500 simulated prices, and $n$ = 1,000 iterations for the Gibbs sampler.

**Table 6: Estimates of *c* using data generated with different levels of correlation of $u_t$**

a) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{0.5}$

| | $\rho = \mathbf{0}$ | | | $\rho = \mathbf{-0.2}$ | | | $\rho = \mathbf{-0.4}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ |
| *RM* | 1.00 | 0.00 | | 1.02 | 0.00 | | 1.05 | 0.01 | |
| *CDP* | 0.89 | 0.01 | | 0.90 | 0.01 | | 0.91 | 0.01 | |
| *TW* | 1.20 | 0.04 | | 1.20 | 0.04 | | 1.20 | 0.04 | |
| *CFTC* | 1.36 | 0.13 | | 1.36 | 0.13 | | 1.35 | 0.13 | |
| *HAS* | 0.98 | 0.00 | -0.10 | 0.98 | 0.00 | -0.25 | 0.98 | 0.00 | -0.38 |
| *ABS* | 0.98 | 0.00 | -0.11 | 0.98 | 0.00 | -0.25 | 0.98 | 0.00 | -0.38 |
| *AVG* | 0.98 | 0.00 | -0.10 | 0.99 | 0.00 | -0.25 | 0.99 | 0.00 | -0.38 |

b) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{1}$

| | $\rho = \mathbf{0}$ | | | $\rho = \mathbf{-0.2}$ | | | $\rho = \mathbf{-0.4}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ |
| *RM* | 1.00 | 0.01 | | 1.10 | 0.01 | | 1.18 | 0.04 | |
| *CDP* | 0.81 | 0.04 | | 0.89 | 0.01 | | 0.97 | 0.00 | |
| *TW* | 1.41 | 0.17 | | 1.41 | 0.17 | | 1.41 | 0.17 | |
| *CFTC* | 1.54 | 0.29 | | 1.56 | 0.32 | | 1.58 | 0.34 | |
| *HAS* | 0.72 | 0.08 | -0.19 | 0.77 | 0.05 | -0.25 | 0.81 | 0.04 | -0.32 |
| *ABS* | 0.86 | 0.02 | -0.20 | 0.88 | 0.02 | -0.26 | 0.89 | 0.01 | -0.33 |
| *AVG* | 0.86 | 0.02 | -0.20 | 0.84 | 0.03 | -0.26 | 0.82 | 0.03 | -0.33 |

c) Parameters of the simulated data: $c = 1$, $p_{buy} = 0.5$, and $\boldsymbol{\sigma_u} = \mathbf{1.5}$

| | $\rho = \mathbf{0}$ | | | $\rho = \mathbf{-0.2}$ | | | $\rho = \mathbf{-0.4}$ | | |
|---|---|---|---|---|---|---|---|---|---|
| | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ | $E[\hat{c}]$ | **MSE** | $E[\hat{\rho}]$ |
| *RM* | 1.00 | 0.01 | | 1.20 | 0.05 | | 1.38 | 0.15 | |
| *CDP* | 0.79 | 0.05 | | 0.97 | 0.01 | | 1.12 | 0.02 | |
| *TW* | 1.66 | 0.44 | | 1.66 | 0.44 | | 1.66 | 0.44 | |
| *CFTC* | 1.78 | 0.61 | | 1.82 | 0.68 | | 1.86 | 0.74 | |
| *HAS* | 0.68 | 0.11 | -0.13 | 0.80 | 0.04 | -0.21 | 0.88 | 0.02 | -0.29 |
| *ABS* | 0.97 | 0.00 | -0.13 | 0.99 | 0.00 | -0.21 | 1.01 | 0.00 | -0.30 |
| *AVG* | 0.97 | 0.00 | -0.13 | 0.99 | 0.00 | -0.21 | 1.01 | 0.00 | -0.30 |

Notes: $\rho$ is the coefficient of correlation of the error term in the Roll model ($u_t$). *RM* (Roll) and *CDP* (Chu, Ding, and Pyun) are serial covariance estimators, *TW* (Thompson and Waller) and *CFTC* (Commodity Futures Trading Commission) are mean absolute price change estimators, *HAS* (Hasbrouck), *ABS* and *AVG* are Gibbs sampler estimators. All estimations come from $k$ = 1,000 simulations using $T$ = 500 simulated prices, and $n$ = 1,000 iterations for the Gibbs sampler.

**Figure 1: Distribution of the half spread when no restriction is imposed**

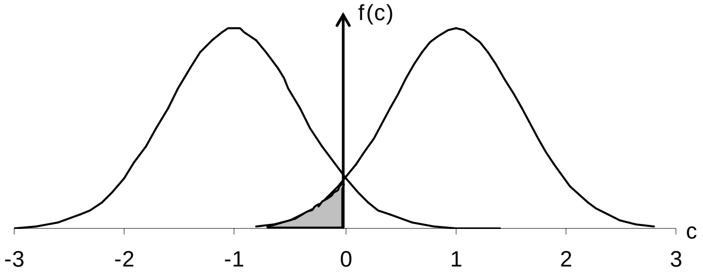


**Figure 2: Price series when $p_{t-1} \approx p_{t+1}$**

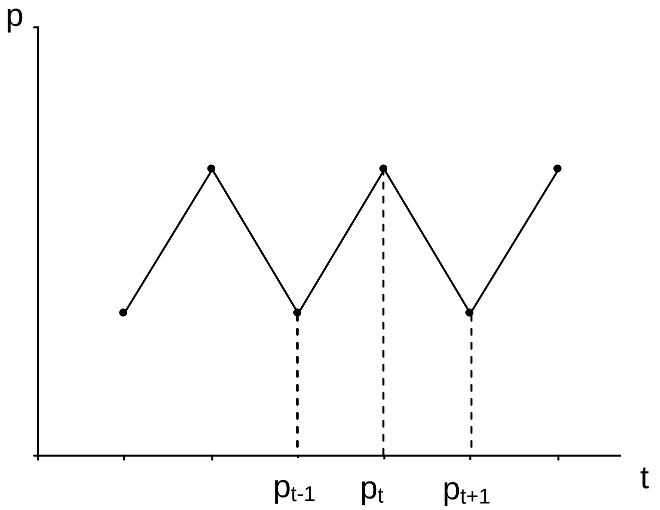
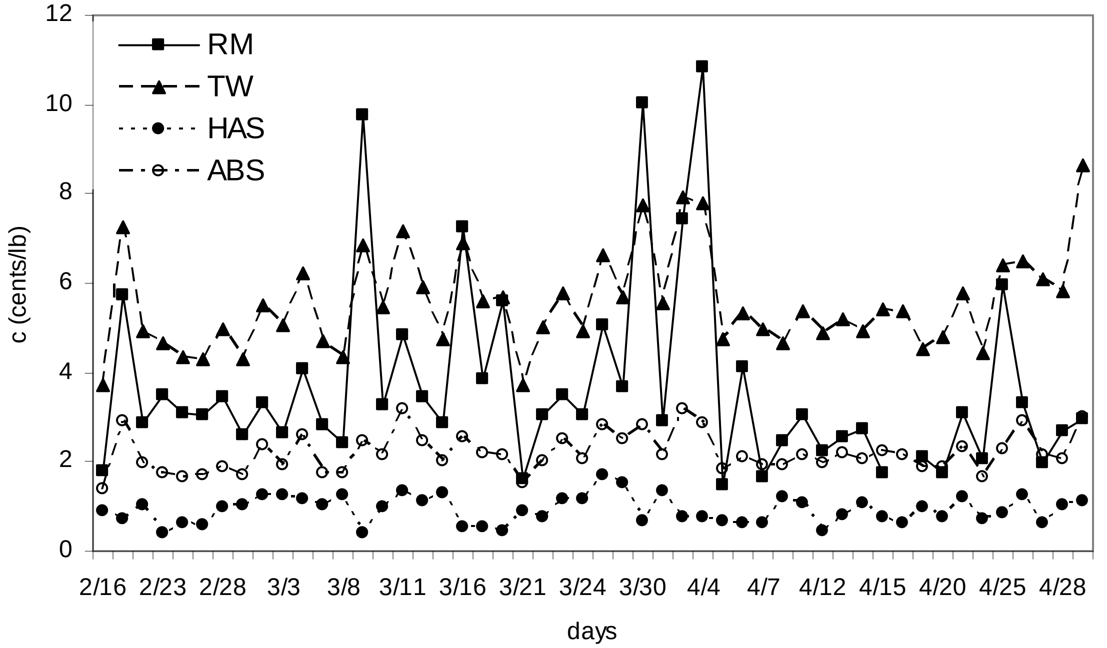
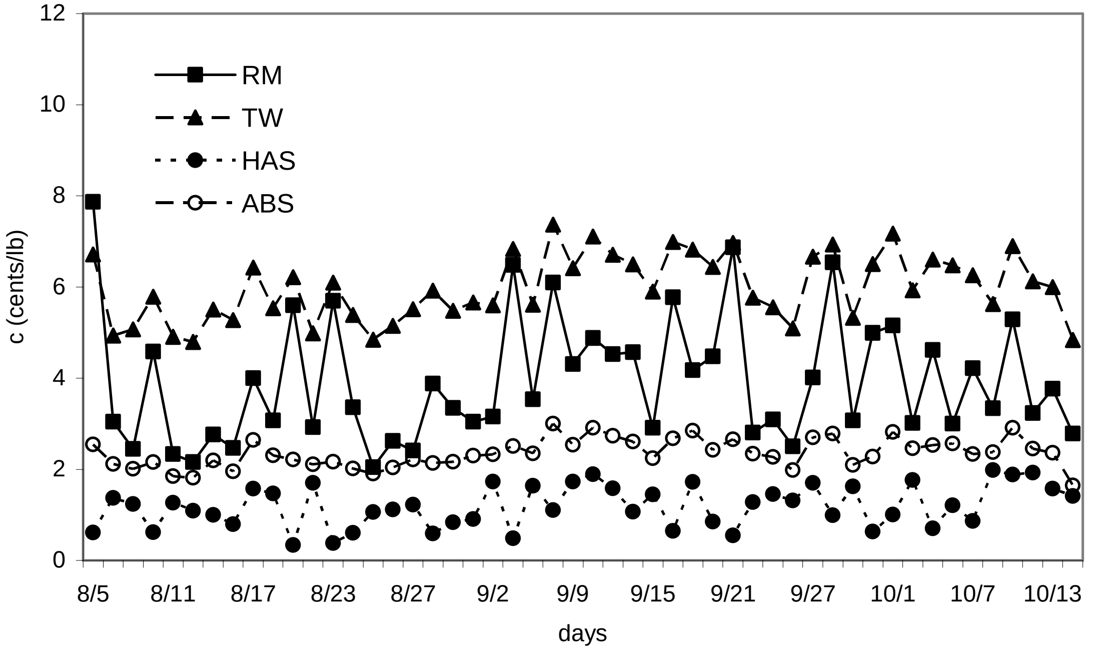


**Figure 3: Bid-Ask spread estimates for the last 50 days of two contracts**

a) April contract for live cattle



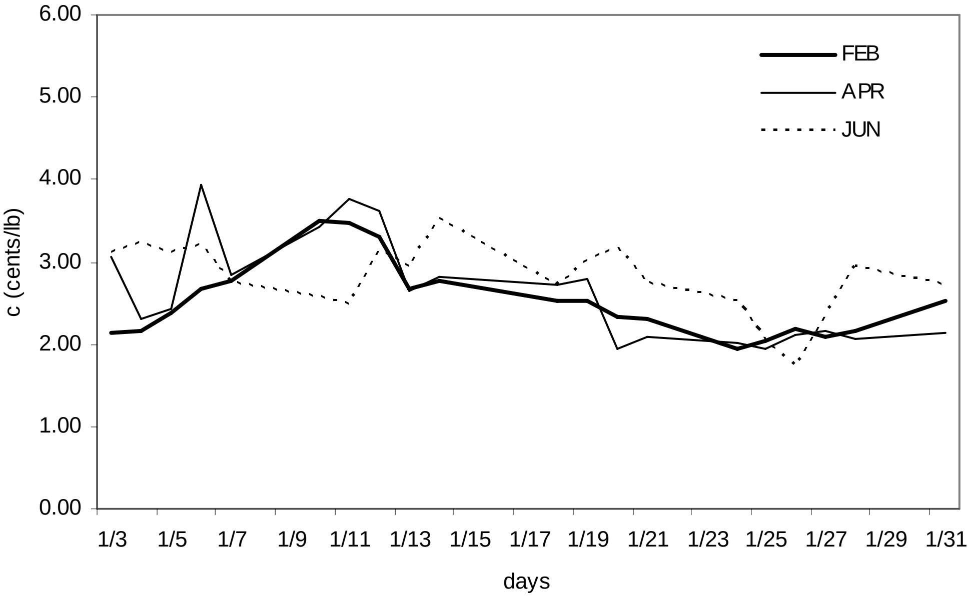
b) October contract for lean hogs



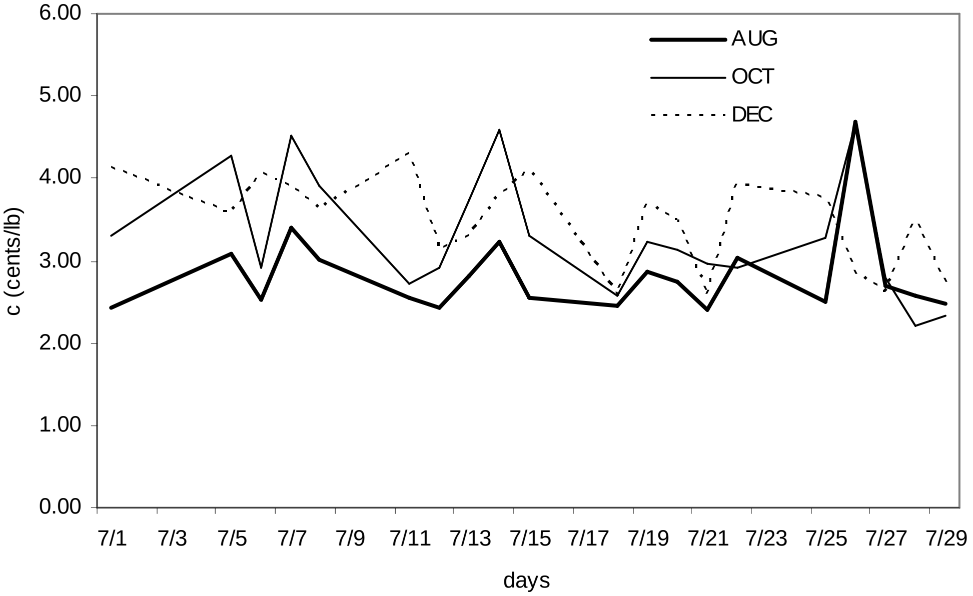
RM (Roll) and is a serial covariance estimator, TW (Thompson and Waller) is a mean absolute price change estimator, and HAS (Hasbrouck) is a Gibbs sampler estimator using a truncated distribution of *c* and ABS is a Gibbs sampler estimator using absolute values of *c*.

**Figure 4: ABS bid-ask spread estimates for one nearby and two distant contracts**

a) Live cattle



b) Lean Hogs



## Appendix

The Roll model states:

$\Delta p_t = u_t + c\Delta q_t$ $\qquad$ $u_t \sim iid\ N(0, \sigma^2_u)$

$q_t$={*-1* for a sell, *+1* for a buy} $\qquad$ $q_t \sim Bernoulli(1/2)$

Then, the possible values for $\Delta q_t$ are:

| $q_{t-1}$ | $q_t$ | $\Delta q_t$ | P($\Delta q_t$) = P($q_t$) P($q_{t-1}$) |
|---|---|---|---|
| -1 | -1 | 0 | ¼ |
| -1 | +1 | 2 | ¼ |
| +1 | -1 | -2 | ¼ |
| +1 | +1 | 0 | ¼ |

Therefore, P($\Delta q_t$=-2) = P($\Delta q_t$=2) = ¼ and P($\Delta q_t$=0) = ½

Let $c$>0 and $f_{\Delta p}(p)$ be the *pdf* for $\Delta p$

$f_{\Delta p}(p) = f_{\Delta p/\Delta q=-2}(p)\ P(\Delta q=-2) + f_{\Delta p/\Delta q=0}(p)\ P(\Delta q=0) + f_{\Delta p/\Delta q=2}(p)\ P(\Delta q=2)$

$\qquad = \frac{1}{4} f_{\Delta p/\Delta q=-2}(p) + \frac{1}{2} f_{\Delta p/\Delta q=0}(p)\ P(\Delta q=0) + \frac{1}{4} f_{\Delta p/\Delta q=2}(p)\ P(\Delta q=2)$

Based on the Roll model,
If $\Delta q = 0 \Rightarrow \Delta p = u \Rightarrow f_{\Delta p/\Delta q=0} \sim N(0, \sigma^2_u)$
If $\Delta q = -2 \Rightarrow \Delta p = u - 2c \Rightarrow f_{\Delta p/\Delta q=-2} \sim N(-2c, \sigma^2_u)$
If $\Delta q = 2 \Rightarrow \Delta p = u + 2c \Rightarrow f_{\Delta p/\Delta q=2} \sim N(2c, \sigma^2_u)$

Therefore,

$$\begin{cases} f_{\Delta p/\Delta q=0}\ (p) = \dfrac{1}{\sqrt{2\pi}\sigma_\mu} e^{-\frac{p^2}{2\sigma_\mu^2}} \\ f_{\Delta p/\Delta q=-2}\ (p) = \dfrac{1}{\sqrt{2\pi}\sigma_\mu} e^{-\frac{(p+2c)^2}{2\sigma_\mu^2}} \\ f_{\Delta p/\Delta q=-2}\ (p) = \dfrac{1}{\sqrt{2\pi}\sigma_\mu} e^{-\frac{(p-2c)^2}{2\sigma_\mu^2}} \end{cases}$$

Note that $f_{\Delta p}(p)$ is symmetric around $p$=0 and $f_{|\Delta p|}(p)$ is twice $f_{\Delta p}(p)$ for $p$>0,

$$f_{|\Delta p|}(p) = \begin{cases} 0 & p < 0 \\ f_{|\Delta p|}(0) = \dfrac{1}{\sqrt{2\pi}\sigma_u}\dfrac{1}{2}\left(1 + e^{\frac{-4c^2}{2\sigma_u^2}}\right) & p = 0 \\ \dfrac{1}{2} f_{\Delta p/\Delta q=-2}(p) + f_{\Delta p/\Delta q=0}(p) + \dfrac{1}{2} f_{\Delta p/\Delta q=2}(p) & p > 0 \end{cases}$$

The mean value of $f_{|\Delta p|}$ *(p)* is:

$$\begin{aligned} E\left[|\Delta p|\right] &= \int_0^\infty p\, f_{|\Delta p|}(p)\; dp \\ &= \frac{1}{2}\left[\frac{\sigma_u}{\sqrt{2\pi}} e^{\frac{-4c^2}{2\sigma_u^2}} - 2c(1 - F(2c))\right] + \frac{\sigma_u}{\sqrt{2\pi}} + \frac{1}{2}\left[\frac{\sigma_u}{\sqrt{2\pi}} e^{\frac{-4c^2}{2\sigma_u^2}} + 2c(1 - F(-2c))\right] \\ &= \frac{\sigma_u}{\sqrt{2\pi}} + \frac{\sigma_u}{\sqrt{2\pi}} e^{\frac{-4c^2}{2\sigma_u^2}} + c\big(F(2c) - F(-2c)\big) \end{aligned}$$

where *F* is the normal cumulative distribution function with mean zero and variance $\sigma_u^2$