# New Classical and Bayesian Estimators for Classifying Trade Direction in the absence of Quotes

**Michael Bowe[a,b], Sungjun Cho[a], Stuart Hyde[a] and Iljin Sung[a]**

## This Draft: May 2018

We propose new methods for estimating the effective bid-ask spread and classifying trading intentions without access to quotes. Our state space approach utilizes both classical and Bayesian estimators. We extend Hasbrouck's (2004) methodology by simultaneously allowing for both unbalanced and autocorrelated order flow, and a role for informational asymmetry. Our methods are easy to implement in practice and provide simple parametric alternatives to both the nonparametric bid-ask spread estimators proposed by Chen, Linton, Schneeberger, and Yi (2016) and the various trade classification algorithms discussed in Easley, Lopez de Prado and O'Hara (2016). For illustrative purposes, we apply our approach to an analysis of the trading patterns in the CME's gold futures contract during a period incorporating uncertainty in financial markets as a result of the UK's 2016 Brexit referendum.

[a] Division of Accounting and Finance, Alliance Manchester Business School, University of Manchester, Booth Street East, Manchester, M13 9SS, UK. email: mike.bowe@mbs.ac.uk, sungjun.cho@mbs.ac.uk, stuart.hyde@mbs.ac.uk, iljin.sung@manchester.ac.uk
[b] Department of Accounting and Finance, University of Vaasa, Wolffintie 34, 65200 Vaasa, Finland.

We wish to acknowledge the comments of Ian Garrett, Julian Williams and Sarah Zhang on earlier drafts of this paper.

## 1. INTRODUCTION

A variety of approaches exist in the literature to compute accurate estimates of trading costs in financial markets in order to evaluate the impact of changing liquidity conditions on market performance. These liquidity proxies are typically computed based on quotes and employ various methods of assigning trading intentions, where a trade is classified as a buy (sell) if the active (i.e., liquidity consuming) side of the trade is a buyer (seller). Trade direction indicators based on such classifications are then used to measure the information content of trades (e.g., Hasbrouck (1991)) and to predict liquidity crashes (e.g., Easley, Lopez de Prado and O'Hara (2012)).

Accurately estimating liquidity often proves to be an elusive task. A major obstacle relates to the stipulations placed on the data, since liquidity estimates often require accurate observations on both intra-daily bid-ask quotes and transaction prices. As a result, trade classification is never straightforward. Historically, the Lee and Ready (1991) algorithm (LR), based on both quote and price changes, is the most popular trade classification algorithm. However, recent literature suggests that with the advent of high frequency trading in markets, the accuracy of LR algorithm is potentially undermined. For example, Easley, Lopez de Prado and O'Hara (2016) argue that in electronic limit order markets, some with order cancellation rates of 98% or more, trade classification algorithms based on proximity to bid and ask quotes are severely compromised. Holden and Jacobsen (2014) using the TAQ data, and Panayides, Shohfi, and Smith (2014) employing Euronext Paris data provide empirical support for these claims. Furthermore, as Holden, Jacobsen, and Subrahmanyam (2015) observe, the trading environment in many financial markets (such as futures and foreign exchange trading) lacks transparency, in the sense actual quotes are not directly observable in the intra-daily data record. Such information must somehow be discerned from the data.

To overcome these data limitations requires developing an empirical methodology to extract a liquidity proxy and classify trading intentions without access to quotes. In this regard, Hasbrouck (2004) develops both a liquidity measure and trade classification algorithms in the absence of quotes using variations of the Roll (1984) model, proposing a new Bayesian approach by assuming an i.i.d. normal distribution for price innovations and latent independent trade indicators. Four representative CME futures contracts illustrate application

of the methodology. Subsequently, Van der Wel, Menkveld, and Sarkar (2009) develop the equivalent classical maximum likelihood estimation (MLE) methods by mapping the Roll model onto the regime switching state space model of Kim and Nelson (1999).

Chen, Linton, Schneeberger, and Yi (2016) document several concerns with the Hasbrouck (2004) approach, including its assumptions of normal price innovations, balanced market order flow, the absence of serial correlation in the trade direction indicators, symmetric information, and constant spreads within the sample period (e.g., for a month). They argue these assumptions could lead to inaccurate estimates, at least during certain trading episodes. These reservations lead Chen, Linton, Schneeberger, and Yi (2016) to propose new nonparametric methods for estimating the bid-ask spread using only transaction prices. Initially, they relax the normality assumption for prices innovations using an empirical characteristics function while maintaining the other assumptions of the Hasbrouck (2004) method. They find that their method produces nearly identical results to the Roll (1984) and Hasbrouck (2004) methods during normal times but performs much better during periods of extreme turbulence. Specifically, analyzing movements in the E-mini futures contract on the S&P 500 during the Flash Crash, they discover that while their estimator is comparable to other methods during most of Flash Crash day, during its peak period, i.e., between 2:45 pm and 2:49 pm ET, their spread estimates seem to provide better approximations. The paper also suggests how their proposed framework can accommodate certain other extensions, such as: unbalanced order flow, serially dependent latent trade indicators, or adverse selection. However, there are several caveats to their approach. First, no empirical analysis is undertaken involving these extensions, possibly reflecting the pervasive curse of dimensionality when applying such nonparametric methods. Second, they develop each extension in isolation, without simultaneously relaxing the limiting features they identify in prior models. Finally, as their focus is on developing new methods to estimate the bid-ask spread, they do not provide a filtering algorithm to obtain the latent trade direction indicators.

The central contribution of this paper is to develop easy-to-implement Bayesian and MLE estimators by extending both Hasbrouck (2004) and Van der Wel, Menkveld, and Sarkar (2009) to simultaneously accommodate several of the omitted features evaluated in Chen, Linton, Schneegerger, and Yi (2016), namely unbalanced and autocorrelated order flow and informational asymmetries.

The second major contribution of this paper is to provide trade direction classification mechanism without recourse to quotes. These classification systems utilise both Bayesian MCMC methods and classical filtering and smoothing algorithms for latent trade direction indicators. Recently, Easley, Lopez de Prado, O'Hara (2016) propose a new conceptual framework for classifying trades, taking the perspective of a Bayesian statistician with priors on the unobservable information (buy or sell indicator), who is trying to extract trading intentions from observable trade data. They compare the strengths and weakness of several rules against an ideal Bayesian rule. We propose that certain familiar structural empirical market microstructure models, such as those we employ in this analysis, provide plausible approximations to their ideal Bayesian trade classification approach. In particular, these models employ a Markov switching process as the underlying process governing the dynamics of the unobservable buy-sell indicator, and treat the measurement equations as a plausible data generating process for the observed data relating to the indicator. Thus, we propose using estimates of the autocorrelated trade direction indicators, or the buy-sell indicator, as the model consistent, trade classification algorithm.

For purposes of illustration, we apply our proposed approach to analyse trading behaviour in the gold futures contract trading on the CME over the two month period from May 2016 to June 2016, a timeframe incorporating the UK Brexit referendum. Specifically, we first estimate the effective spread, and subsequently decompose it into non-informational and informational components, computing daily correlation estimates of classified trades between our model-consistent trade classification rules and those we obtain from the Tick rule.

The main findings are as follows. First, we obtain almost identical results from both classical MLE and Bayesian methods in all empirical models throughout the sample period. Second, we find estimates of daily trade direction indicators to be highly autocorrelated, leading to measured bid-ask spreads being larger, in an economically meaningful sense, than those obtained from alternative estimates employing independent trade direction indicators. Third, we find strong statistical support for asymmetric information models of the type proposed by Glosten and Harris (1988) in the presence of latent and autocorrelated trade direction indicators. The results provide evidence that the trade impact coefficients implied by the asymmetric information model, which reflect Kyle's lambda, are important elements of liquidity. Fourth, when comparing the Roll model and Tick rule we find that the daily

correlation estimates of the classified trades are almost always above 0.99, indicating that the trade classification we obtain from the Roll model used in Hasbrouck (2004) and the Tick rule are essentially identical. Finally, our model consistent trade classification algorithm based on an extended GH model provides very similar results to the Tick rule during normal trading periods. However, in the presence of greater uncertainty when trading potentially generates a greater price impact (relating from to order flow imbalances), our trade classification indicator often diverges significantly from those we obtain using the Tick rule.

As Easley, Lopez de Prado, and O'Hara (2016) maintain that Tick rule classifications appear particularly problematic in periods of high volatility exhibiting imbalances in order flow, we believe the approach to trade classification we propose shows some promise. Importantly, we do not claim that our trade classification system is superior to other rules. As Easley, Lopez de Prado, and O'Hara (2016) note, each trade classification rule may demonstrate both strengths and weakness, depending on the underlying market characteristics. Instead, we maintain that our approach may be best suited to classifying trades consistently in environments where a variant of state space models with regime switching yields a realistic approximation to the trading conditions. Moreover, our methods have the advantage of providing easy-to-implement model consistent trade classification algorithms using both Bayesian and Classical estimation methods. As such, we believe they may be a useful addition to the empirical microstructure tool kit.

The remainder of this paper is organized as follows. Section 2 briefly reviews the Roll (1984) structural market microstructure model and its subsequent generalizations leading to richer information-based models. Our focus here is on resolving estimation issues linked to the model parameters. Section 3 presents the classical and Bayesian estimation methods we propose. We outline data sources and present and discuss the main empirical results in section 4. Section 5 concludes the paper.

## 2. Empirical Structural Market Microstructure Models

### 2.1. The Roll model

The Roll (1984) model is a parsimonious structural market microstructure model of the bid-ask spread. The model decomposes the dynamics of the asset pricing process into two components, namely: (i) changes in the "efficient price" reflecting the fundamental value of the security conditional on all publicly available information, and (ii) the costs associated with the trading process. This model is initially derived by assuming a competitive dealer market with fixed transaction costs and symmetric information, in which dealers set their bid-ask quotes to recover their costs of making a market. However, in modern financial markets, high frequency trading firms typically act as market makers, by placing passive orders at various levels of the order book to earn tiny margins on large bets (Easley, Lopez de Prado, O'Hara (2012)). Indeed, Hendershott and Menkveld (2014) propose a more general definition of liquidity suppliers (market makers) in modern financial markets as agents who trade against price pressures. This interpretation of the market maker is consistent with the Roll model, where the market maker buys at a discount (negative price pressure) and sells at a premium (positive price pressure).

The price dynamics in the Roll model can be represented as follows. Denote the efficient price by $M_t$ with log ($M_t$) = $m_t$ and the transaction price by $P_t$ with log ($P_t$) = $p$. The evolution of these two prices can be depicted as:

$$p_t = m_t + cq_t, \quad m_t = m_{t-1} + u_t, u_t \sim N(0, \sigma_u^2) \qquad (1)$$

where $q_t$ is a regime switching variable with $q_t \in \{-1, +1\}$ and

$\Pr[q_t = 1 \mid q_{t-1} = 1] = 0.5, \Pr[q_t = -1 \mid q_{t-1} = -1] = 0.5$.

The Roll model contains two sources of randomness. It assumes the efficient price evolves as a random walk, with the i.i.d. innovation term ($u_t$) reflecting public information. The trade direction indicator $q_t$ is a random variable taking one of two values, +1 (-1) for a buyer (seller) initiated trade. Buyer and seller initiated trades are assumed to be equally probable, and in the Roll model $q_t$ is independent of $u_t$, so the direction of trade is independent of

changes in the efficient price. This effectively eliminates any influence of asymmetric information in the model, and it is one of the key assumptions we relax later in the paper. The term *c* is interpreted as the (log of) the effective execution cost paid by an active buyer or seller. The Roll specification implies:

$$\Delta p_t = m_t + cq_t - m_{t-1} - cq_{t-1} = c\Delta q_t + u_t, \qquad (2)$$

Roll proposes a moment estimator to compute estimates of bid-ask spreads based only on transaction prices. However, Roll's estimate is feasible only if the first-order sample autocovariance is negative and as a result Roll's reported spread estimator is often biased downward. To reduce this downward bias, Hasbrouck (2004) proposes a Bayesian method to estimate model parameters, assuming normal distributions characterise the innovation term ($u_t$) and latent independent trade indicators. Van der Wel, Menkveld, and Sarkar (2009) develop alternative classical MLE methods for the Hasbrouck (2004) model , and Chen, Linton, Schneeberger, and Yi (2016) use nonparametric methods to relax the normality assumption. . However, Chen, Linton, Schneeberger, and Yi (2016) also point out several remaining problems with these econometric approaches, such as the assumptions of balanced market order flow, symmetric information and the absence of serial correlation in the trade direction indicators. In the following section we proceed to relax these assumptions and provide an extension to the econometric methods developed by Hasbrouck (2004) and Van der Wel, Menkveld, and Sarkar (2009).

## 2.2 Generalizations of the Roll model

**(i) Autocorrelation in order arrival and unbalanced market order flow**

Choi, Salandro and Shastri (1988) provide several reasons, such as information disclosure concerns leading to strategic trading behaviour (order fragmentation), for the existence of serially correlated trade arrival in financial markets, and extend the Roll model to incorporate autocorrelated trade direction indicators. In this paper we use the following model (henceforth the extended Roll model (MS)) to accommodate these stylised facts.

$$p_t = m_t + cq_t, m_t = m_{t-1} + u_t, u_t \sim N\left(0, \sigma_u^2\right) \qquad (3)$$

where the trade direction indicator, $q_t$ is a random variable taking values of 1 or -1, governed by the following Markov process:

$$\Pr[q_t = 1 \mid q_{t-1} = 1] = P \quad \text{and} \quad \Pr[q_t = -1 \mid q_{t-1} = -1] = Q$$

**(ii) Adverse selection**

Easley, Lopez de Prado, and O'Hara (2012, p.1457) define adverse selection in modern limit order markets as the "natural tendency for passive orders to fill quickly when they should fill slowly and fill slowly (or not at all) when they should fill quickly". They also explain that such a definition is consistent with market microstructure models proposed by Glosten and Milgrom (1985) and Kyle (1985). In these models, order flow is informative for subsequent price moves as it reflects the level of informed trading. These models provide insights into the behaviour of market participants in the presence of informational asymmetries and motivate further modifications to the original Roll model. These modifications allow the efficient price to be at least partially driven by the trade direction indicator variable, capturing a key feature of asymmetric information microstructure models, namely that trade characteristics may convey information correlated with a trader's private information.

To accommodate the above stylised features of trading, we propose a model extension which incorporates both a term capturing latent and autocorrelated trade direction indicators and an additional one reflecting potential adverse selection costs.. In this model, only the unexpected component of the trade direction indicator series produces any effect on the efficient price. We represent this model (henceforth the extended GH model (GH)) as follows:

$$p_t = m_t + cq_t, m_t = m_{t-1} + (\lambda_0 + \lambda_1\sqrt{V_t})\left(q_t - E\left[q_t \mid \psi_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right) \qquad (4)$$

where $\lambda_0$ and $\lambda_1$ represent that fixed and variable permanent price impact costs, respectively, and $\psi_{t-1}$ denotes the available information set, up to time t-1. The trade direction indicator, $q_t$, follows the Markov process in (3). Following a straightforward reformulation, we can express this process as the autoregressive (AR) process.

$$q_{t+1} = (P - Q) + (P + Q - 1)\,q_t + \varepsilon_{t+1},\ \mathrm{E}[\varepsilon_{t+1} \mid q_t] = 0 \qquad (5)$$

This formulation of the AR(1) process for the trade direction indicators is used previously in the literature on several occasions (e.g., Madhavan, Richardson, and Roomans (1997)). In summary, in our proposed model, $E\left[q_t \mid \psi_{t-1}\right]$ can be expressed as $(P-Q)+(P+Q-1)q_t$, and in its final form, the empirical model we estimate may be depicted as follows:

$$p_t = m_t + cq_t,$$
$$m_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \mu - \rho q_{t-1}\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right) \quad (6)$$

where $q_t = \{1,-1\}, \Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P$, $\Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\mu = P - Q$, $\rho = P + Q - 1$.

The model has a reduced form representation, given by:

$$\Delta p_t = c\Delta q_t + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \mu - \rho q_{t-1}\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right) \quad (6')$$

In this framework, the market-maker requires compensation not only for the costs of processing an order ($c$) as in the Roll model, but also for the adverse selection risk of supplying liquidity to an informed trader, where $\lambda_1\sqrt{V_t}$ captures the adverse selection component of a trade's price impact. The literature has a variety of interpretations of the coefficient $\lambda_1$ from a measure of Kyle's lambda, the slope of the price impact curve arising from asymmetric information effects, to the (inverse) market depth parameter (Brennan and Subrahmanyam (1996)). Prior to proceeding, we note that Chen, Linton, Schneeberger, and Yi (2016) provide several theoretical extensions of their model, selectively incorporating unbalanced order flow, serially dependent latent trade indicators, and adverse selection. However, in contrast to the present model formulation, no attempt is made to simultaneously accommodate these features, and they do not conduct any extensive empirical analyses.

## 3. Estimation Methods for Structural Market Microstructure Models

Van der Wel, Menkveld, and Sarkar (2009) show that the Roll model used in Hasbrouck (2004) can be interpreted as a state space model. For example, one natural interpretation of equation (3) is a state space system with measurement and transition equations as follows:

Measurement equation: $p_t = m_t + cq_t, q_t \in \{-1,+1\}$

Transition equation: $m_t = m_{t-1} + u_t, u_t \sim N(0, \sigma_u^2)$

Employing the same reasoning, we can reformulate the extended GH model with asymmetric information (equation (6)) as a state space model, with associated equations given by:

Measurement equation: $p_t = m_t + cq_t$

Transition equation: $m_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \left((P-Q) + (P+Q-1)q_{t-1}\right)\right) + u_t$

Recently, many papers adopt similar state space formulations to estimate market microstructure models. For example, Menkveld, Koopman, and Lucas (2007) model a high-frequency price series as the sum of efficient price series, reflecting permanent price effects as above, and stationary series capturing transitory price effects. A similar framework also is subsequently adopted in Menkveld (2013), Brogaard, Hendershott, and Riordan (2014), and Hendershott and Menkveld (2014).

### 3.1. Bayesian Markov chain Monte Carlo (MCMC) methods

Hasbrouck (2004) develops a Bayesian Gibbs sampling approach to estimating the Roll model with normality assumption governing price innovations (equation (3)), in which the parameters ($c$ and $\sigma_u$) are considered to be random variables (reflecting the statistician's uncertainty). He motivates the technique's adoption with reference to its ability to accommodate important economically meaningful latent data such as trade direction indicators ($q = \{q_1, q_2, \ldots, q_T\}$), which are "suppressed in the GMM estimation" (Hasbrouck 2004, p.311). Inference is based upon a series of transaction prices through time: $p = \{p_1, p_2, \ldots, p_T\}$, with knowledge of $q = \{q_1, q_2, \ldots, q_T\}$ and $p$ sufficient to determine the efficient price, $m = \{m_1, m_2, \ldots, m_T\}$. The respective joint distribution function $F(c, \sigma_u, q \mid p)$, summarizes the full posterior over parameters and latent data. To estimate the present model, we use Hasbrouck's MatLab codes available in his website (http://people.stern.nyu.edu/jhasbrou/).

In this paper, we extend the Roll model of Hasbrouck (2004) by adding unbalanced order flow and autocorrelated trade direction indicators. Our extension (equation (4)) has two extra

transition probability parameters ($P$ and $Q$) and $F(c,\sigma_u,q,P,Q\,|\,p)$ summarizes the full posterior over parameters and latent data. While there is no tractable closed-form representation for this joint distribution function, $F(c,\sigma_u,q,P,Q\,|\,p)$, the full conditional (posterior) distributions for the parameters are often tractable. In estimating both equations (3) and (4), for example, conditional on $q$, the equation $\Delta p_t = c\Delta q_t + u_t$ can be treated as a simple normal linear regression specification in which $c$ and $\sigma_u$ are the regression coefficient and residual standard deviation, respectively. It is also possible to express the conditional distribution of $q$ based on the model parameters and data. The major differences between Hasbrouck's (2004) Gibbs sampling algorithm and the Gibbs sampling method we adopt to estimate the model in equation (4) arises in relation to the simulation of $q_t$.[1]

The power of Bayesian analysis using the Gibbs sampler is that it requires only the conditional distributions to numerically recover the joint distribution function. The Gibbs sampler is an iterative procedure of drawing each parameter, or latent trade direction indicator, sequentially. Initially, the parameters and latent trade direction indicators are set equal to some arbitrary values $\{c^{(0)},\sigma_u^{(0)},q^{(0)},P^{(0)},Q^{(0)}\}$, although efficient estimation typically requires specifying reasonable starting values, such as GMM estimates of ($c$ and $\sigma_u$), whenever they are available. In the subsequent iterations, all parameters and latent trade direction indicators except for the component being drawn are taken as given, and each component is updated sequentially. For example, the second iteration starts with a draw of $c^{(2)}$ conditional on $\left(\sigma_u^{(1)},q^{(1)},P^{(1)},Q^{(1)}\right)$. By repeating this procedure we generate a sequence of draws of unknowns for $j=1,\ldots,n$. The Gibbs principle demonstrates that the limiting distribution of the $n^{\text{th}}$ draw after burn-in samples (as $n\to\infty$) is $F(c,\sigma_u,q,P,Q\,|\,p)$, the desired posterior, and the limiting draw for any parameter is distributed as the corresponding marginal posterior. For example, the limiting density of $c^{(n)}$ is $f(c\,|\,p)$. The number of simulation must be sufficiently large so that dependence on the initial conditions becomes insignificantly small.

[1] We explain the details of the Gibbs sampling algorithms for these two methods and the differences between them in Appendices A1 and A2.

In any MCMC estimation, it is crucial to ensure convergence of the chain in order to undertake correct statistical inference. In this paper, as a diagnostic to check for the convergence of our MCMC algorithm, we compute and report the effective sample size based on the inefficiency factors and the p-value of Geweke's (1992) convergence diagnostics test for the model parameters.[2] Finally, based on the simulation outputs, we estimate population parameters of the posterior using standard time series analysis techniques, noting that the sample mean of $c^{(j)}$ is a consistent estimate of $E[c \mid p]$ and the sample variance is a consistent estimate of $Var[c \mid p]$. We can also interpret trade direction indicators we estimate as outputs from a model-consistent trade classification algorithm.

To estimate the most complex model which incorporates adverse selection (equation (6)), we need to develop a new Bayesian Algorithm to simulate iteratively $F(c, \lambda_0, \lambda_1, \sigma_u, P, Q, q \mid p)$ .This is for the following reasons. First, the innovation in autocorrelated $q_t$ impacts the efficient price. Second, this model involves two extra parameters capturing the price impact of trades ( $\lambda_0$ , $\lambda_1$ ), while the two transition probabilities ( $P$ and $Q$ ) also appear in the regression specification. The implication is that we cannot use the customary Gibbs sampling estimation algorithm to estimate these parameters. In response, we develop a tailored random walk Metropolis Hastings algorithm to undertake this task.[3]

### 3.2 Classical MLE methods

Van der Wel, Menkveld, and Sarkar (2009) develop alternative MLE methods for Hasbrouck's (2004) Roll model formulation (equation (3)). In this paper, we extend their approach in two ways. First we incorporate order flow imbalances and autocorrelated trade direction indicators (equation (4)), and second, we include an adverse selection term (equation (6)). We can interpret these models as state space models with regime switching, as explained in the previous section. In ordinary state space models with normally distributed

[2] We provide more details in Appendix B.

[3] We present the details in Appendix A3.

shocks, a Kalman filtering technique is employed to construct the likelihood function utilising prediction error decomposition. However, several models used in this paper incorporate a discrete regime switching variable ($q_t$). In such cases, Kim and Nelson (1999) demonstrate we can estimate the model with a composite filter, which is a combination of a Kalman and a Hamilton filter. Based on this nonlinear filter we obtain an approximate likelihood function, and utilise MLE methods for estimation.[4] One critical issue with this approach is that it is difficult to precisely quantify the bias caused by using this approximate filter. Fortunately, we can avoid undertaking the approximation by expressing several models in their reduced form. For example, as previously shown (equation (6')), the extended GH model with asymmetric information can be written as:

$$\Delta p_t = c\Delta q_t + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \left((P-Q) + (P+Q-1)q_{t-1}\right)\right) + u_t$$

On this basis, we can interpret this version of the extended GH model as a standard regime switching model and use a Hamilton filter to construct likelihood function values.[5]

### 3.3 Discussion of the estimated trade direction indicator, $q_t$

Our core proposal is that the autocorrelated trade direction indicators, $q_t$, we estimate from the previous models can be considered to be model consistent trade classification algorithms. Specifically, we can directly recover the trade direction indicator, $q_t$, from the outputs of Bayesian estimation, while for classical MLE methods, we can employ filtering and smoothing algorithms to compute the probability of $q_t$ for each trade. In other words, once we estimate the empirical market microstructure models we present in this paper, we can determine whether each trade is initiated by a buyer ($q_t = 1$) or a seller ($q_t = -1$).

We note earlier that the advent of high frequency trading platforms calls into question the accuracy of traditional trade classification systems such as the Lee and Ready (1991) algorithm (LR In lieu, Easley, Lopez de Prado, O'Hara (2016) propose a new conceptual

[4] Chapter 5 of Kim and Nelson (1999) provides more details.

[5] We provide precise details of how we implement MLE methods for estimation in Appendix C.

framework for classifying trades. They adopt the perspective of Bayesian statisticians with priors on the unobservable information (here $q_t$), who are trying to extract trading intentions from observable trading data. Ideally, we would like to specify the data generating processes for both the underlying unobservable variables and subsequently for the observed data, conditional on the realizations of the underlying unobservable data. Formulating such specifications may prove a daunting task, and computing closed-form solutions for conditional probabilities is likely to be complex, even when such solutions exist. They claim that every trade classification algorithm can be regarded as an approximation to this Bayesian approach, and that their bulk volume classification (BVC) methodology is conceptually closer to this ideal than traditional approaches such as the Tick rule, since BVC assigns a probability to a given trade being either a buy or sell.

We believe that the empirical market microstructure models we outline in this paper provide another plausible approximation to the ideal Bayesian trade classification approach. For example, in relation to equations (3) and (6), we can interpret the relevant transition equations as the data generating process for the underlying unobservable variables, $q_t$, and the measurement equations as the plausible data generating process for the observed data relating to $q_t$. While the extent to which these empirical market microstructure models capture market reality remains unclear, much research employs these models as the empirical basis for their investigations in this area.[6]

Note, we do not claim that our trade classification system is generally superior to other existing rules such as the Tick and BVC rules. As Easley, Lopez de Prado and O'Hara (2016) maintain, each trade classification rule may have advantages which are only manifest in differing trading environments: with less noisy data the Tick rule may prove to be generally superior to the BVC, while with noisy data the BVC may prevail. We believe that our approach is better suited to situations where a variant of state space models incorporating regime switches better approximates the dynamics of the trading environment. In such situations, our proposed methods provide model consistent trade classification algorithms using both Bayesian and Classical methods which are easy to implement in practice.

---

[6] Hasbrouck (2007) provides a comparative summary of relevant literature.

## 4. EMPIRICAL ANALYSIS

### 4.1 Data decription

We conduct the empirical implementation of our proposed trade classification methods using data from gold futures trading on the Chicago Mercantile Exchange (CME) during May and June 2016. We select this particular asset and time period for the following reasons. First, gold futures (ticker symbol GC), are among the most widely traded of all futures contracts worldwide, and gold is often considered a "safe haven" asset at times of global economic uncertainty, such as the period surrounding the UK's Brexit referendum on June 23, 2016 which we deliberately include in our sample for precisely this reason. Second, as Easley, Lopez de Prado, and O'Hara (2016) explain, the gold futures market is less fragmented than its spot market. Each contract trades on a single market, and trading data is less noisy, since all trades are mandated to occur at either the best bid or the best offer and trades between the spread are not permitted. Further, while New York futures volume is less than a tenth of the London spot volume, the futures contract plays the key role in the process of price discovery, leading the spot market in incorporating new, gold price-relevant information into asset values, (Hauptifleisch, Putnin, and Lucey (2016)).

Specifically, we select our sample data from the gold futures contract trading on CME's Globex electronic trading platform during the period from May 1, 2016 to June 30, 2016. Electronic trading on CME Globex is available virtually 24 hours a day from Sunday 6:00 p.m. through to Friday 5:00 p.m. Eastern Standard Time (EST)[7], with only a 60-minute break each day beginning at 5:00 p.m. EST[8]. For the empirical analysis, with reference to trading volume, we identify the most actively traded gold futures contract on any given day to construct a continuous series. The most active gold futures contract from May 1 to May 26 is the June contract deliverable on any business day in June 2016 (GCM6). On May 27 volume shifts to another contract deliverable during August 2016 (GCQ6). The transaction price series and trading volume data, time-stamped to the microsecond, are sourced from the

[7] Thus, the trading activity on any given date starts at 6:00 p.m. and finishes at 5:00 p.m. EST the following day. For example, on May 1 trading starts at 6:00 p.m. on May 1 and finishes at 5:00 p.m. on May 2.

[8] There is one exception during our sample period: trading on Globex halted at 1:00 p.m. and reopened at 6:00 p.m. on Monday, May 30, 2016 because of the Memorial Day holiday trading schedule.

Thomson Reuter Tick History (TRTH) database. When several orders are time-stamped to the same microsecond they are aggregated to obtain volume-weighted prices and total volumes. Finally, following Marshall, Nguyen, and Visaltanachoti (2012), we winsorize all return variables at the 0.5% and 99.5% levels so data errors are not driving the results.[9]

Table 1 describes relevant features of the gold futures contracts. Each gold futures contract represents 100 troy ounces and is quoted in US dollars per troy ounce. Minimum tick size on the contract is 10 cents per troy ounce. As a proportion of the contract price, average tick size is 0.008% and the standard deviation of the price change is approximately 7 ticks after winsorization at the 0.5% and 99.5% levels. The contracts evidence a fast pace of trading activity, with an average 1.6 seconds between trades. Figure 1 illustrates the daily series of average prices and trading volumes during the sample period.

### 4.2 The Roll model

As our benchmark, we estimate the Roll model each day during our sample period, using both Hasbrouck's (2004) Bayesian, and Van der Wel, Menkveld and Sarkar's (2009) MLE models. To clarify notation and facilitate ensuing discussions, we re-state the Roll model as:

$$p_t = m_t + cq_t, \quad m_t = m_{t-1} + u_t, u_t \sim N(0, \sigma_u^2)$$

where $q_t$ is a regime switching variable with $q_t \in \{-1, +1\}$ and

$$\Pr[q_t = 1 \mid q_{t-1} = 1] = 0.5, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = 0.5 \quad .$$

All model parameters are assumed to be constant during any given day, but can change across days. To guarantee accurate statistical inference, we implement the Bayesian estimations with a 100,000 iteration burn-in period and 250,000 total iterations, increasing these numbers by 10,000 until we simultaneously satisfy both the convergence criteria for the inefficiency factors and Geweke's (1992)'s diagnostic tests for all parameters[10].

[9] Marshall, Nguyen, and Visaltanachoti (2012) clean the data in two steps. They first compute the 5%-trimmed sample mean and standard deviation for each high-frequency liquidity measure, meaning the top and bottom.5% observations are excluded from the trimmed mean and standard deviation calculations. Then they remove observations that are outside the trimmed mean by +/- three standard deviations.

[10] Refer to Appendix B for more details on the inefficiency factors and the Geweke (1992) diagnostic tests. Table A.1 (the column 'Roll (Bayesian)) indicates that convergence criteria are achieved for the Roll model, since the effective size exceeds 1,000 and Geweke's (1992) p-values are greater than 0.05 for all parameters.

**[Table 2 in here]**

Table 2 reveals that we obtain benchmark MLE estimates very close to the Bayesian estimates, validating the use of MLE methods in estimating the model. Computing the daily differentials between the estimates of $c$ and $\sigma_u$ we obtain from the two methods, the average value is zeros and the maximum of the absolute differential values is minuscule (0.1 x $10^{-7}$)[11], confirming the methods provide almost identical estimates. The averages of the percentage effective half spread, $c$, and the standard deviations, $\sigma_u$, are 0.26 x $10^{-4}$ and 0.43 x $10^{-4}$, respectively. Noteworthy is the fact that on the day of the UK's Brexit referendum(June 23), the estimate of $c$ is 0.33 x $10^{-4}$, one of the largest values in the sample, while that for the standard deviation, $\sigma_u$, is 0.96 x $10^{-4}$, an increase of close to 90% when compared to its value the previous day.

### 4.3 The Roll model with autocorrelated $q_t$ : a simulation exercise

We now proceed to introduce autocorrelated trade direction indicators into the analysis, by re-stating the extended Roll model as:

$$p_t = m_t + cq_t, \quad m_t = m_{t-1} + u_t, u_t \sim N(0, \sigma_u^2)$$

where $q_t$ is a regime switching variable with $q_t \in \{-1, +1\}$, and to capture autocorrelation in the trade direction indicators we specify:

$$\Pr[q_t = 1 \mid q_{t-1} = 1] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$$

Simulations are undertaken both to validate the computational accuracy of our estimations and to compare the results we obtain using alternative approaches to estimation. To establish a sound basis for a plausible data generating process in the simulations, we generate the data employing the averages of the parameter estimates we report in Table 4, on the assumption that the trade direction indicator is autocorrelated and has persistent regimes. To minimize simulation errors, for each model we generate the data 50 times, and estimate the parameters and trade direction indicators for each data sample, enabling us to compute simulated sample averages. We then compare the estimation results from the MLE approach with those from our proposed Bayesian Gibbs sampling methods.

[11] Table A.2 provides further details..

**[Table 3 in here]**

Table 3 presents the results from these simulations. The estimates of the standard deviation of changes in the (log) efficient price and the (log) half-spread are $\sigma_u$ and $c$, respectively. Finally, *P* and *Q* are the transition probabilities of the latent regime switching process for the trade direction indicators. The results we label "Bayesian" derive from the Bayesian Gibbs sampler in which trade direction indicators are conditionally simulated and autocorrelated. The results are based on 12,000 sweeps of the sampler, with the first 2,000 discarded. We obtain the alternative estimates MLE estimates applying classical MLE methods to the regime switching models. Once again, we conclude that both methods provide estimates close to the true parameter values we use in the simulations.

### 4.4 The Roll model with autocorrelated $q_t$: gold futures contract estimation

We believe the outcome of the simulation exercises in section 4.3 enable us to confirm the integrity of our chosen estimation methodology, so , using the CME gold futures contract data we now proceed to generate daily estimates of the Roll model incorporating autocorrelated trade direction indicators. As before, we assume constant parameters for this model during any given day, although parameters can vary across days. The Bayesian estimations follow an identical approach to simulation as described in section 4.2.[12] We compare results from the MLE method based on Regime Switching (RS) and our proposed extended MCMC method.

**[Table 4 in here]**

Table 4 presents our estimation results, and once again reveals that the MLE and Bayesian estimates are very similar. First, the average value of the differentials between the daily estimates of the coefficient values of $c$ and $\sigma_u$ are close to zero, and the maximum of the absolute differential values are minuscule ($0.2 \times 10^{-7}$). We also note very small differences in the sets of transition probability estimates[13]. Second, the values of both $c$ and $\sigma_u$ remain

[12] Refer to Appendix B for more details on the inefficiency factors and the Geweke (1992) diagnostic tests. Table A.1 (the column 'Roll (Bayesian)) indicates that convergence criteria are achieved for the Roll model, since the effective size exceeds 1,000 and Geweke's (1992) p-values are greater than 0.05 for all parameters.

[13] Table A.2 provides more details.

significant across days, and the sample averages of the percentage effective half spread $c$, and of standard deviations $\sigma_u$, are 0.43 x $10^{-4}$ and 0.35 x $10^{-4}$, respectively. The transition probabilities for the trade direction indicators are autocorrelated with coefficient values close to 0.7. These empirical results indicate the presence of moderately persistent trade direction indicator regimes. It is interesting to note that on the Brexit referendum day (23 June 2016), the estimate for $c$ is a1.12 x $10^{-4}$, which represents the largest single value in the sample, standard deviation estimate is 0.78 x $10^{-4}$, reflecting an increase of close to 90% in comparison to the previous day. Finally, the persistence measure for the regime ($P+Q-1$) is 0.55, which is much larger than the average estimate of 0.4.

Attention is drawn to another feature of the Brexit referendum day estimates. Both the MLE and Bayesian methods generate substantially larger *c* estimates using the formulation of the Roll model incorporating autocorrelated $q_t$ in comparison to those we derive from our benchmark Roll model (Table 2). One potential explanation helping to justify the relative size of these *c* coefficient estimates (with and without autocorrelated $q_t$) is apparent following consideration of the nature of the GMM estimates we obtain following Roll (1984). Specifically, consider the following moment condition we derive from the autoregressive form of the regime switching process: $q_{t+1} = (P-Q) + (P+Q-1)q_t + \varepsilon_{t+1}$.

The GMM estimates of *c* in this model are given by:

$$\gamma_1 = \mathrm{cov}(\Delta p_t, \Delta p_{t-1}) = E[c\Delta q_t + u_t][c\Delta q_{t-1} + u_{t-1}] = c^2 E[\Delta q_t \Delta q_{t-1}] = c^2\left((P-Q)^2 - (1-P-Q+1)^2\right)$$

implying:

$$c = \sqrt{\frac{-\gamma_1}{(1-P-Q+1)^2 - (P-Q)^2}}$$

As the Roll model assumes $P = Q = 0.5$, the resulting GMM estimate of *c* is: $c = \sqrt{-r_1} = \sqrt{-\mathrm{cov}(\Delta p_t, \Delta p_{t-1})}$.

The critical point to note is that once we account for autocorrelation in the trade direction indicators, the transition probabilities for both buy and sell orders enter into the determination of the value of $c$. From the formula, the fact we experience much more persistent regimes on

the Brexit referendum day justifies the substantially larger (log) half-spread estimates we obtain for *c*.

### 4.5 The extended Glosten and Harris model with autocorrelated $q_t$ : a simulation exercise

We are now in a position to apply our proposed estimation procedures in an asymmetric information setting which allows for informed trading activity to have a permanent impact on the efficient price. Our model incorporates an adverse selection cost component of the effective bid-ask spread which in the present specification is a function of the square root of the signed trade volume $q_t\sqrt{V_t}$ [14]. This reflects existing evidence that larger orders are more likely to contain information which has a permanent effect on the efficient price. Moreover, our estimates of the asymmetric information model assume latent and autocorrelated trade direction indicators, so only innovations in $q_t$ impact the efficient price. This differs from the Hasbrouck (2004) formulation which incorporates independent $q_t$, implying the entire signed order flow impacts the efficient price.

Once again, to facilitate explanations and to clarify notation, we re-state this model:

$$p_t = m_t + cq_t, \quad m_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - E_{t-1}q_t\right) + u_t, u_t \sim N(0, \sigma_u^2)$$

where $q_t$ is a regime switching variable with $q_t \in \{-1, +1\}$ and

$$\Pr[q_t = 1 \mid q_{t-1} = 1] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$$

We validate the computational integrity of our estimation algorithms, and undertake a comparison of the results from our proposed estimation method using simulations. As before, to establish a sound basis for a plausible data generating process in the simulations, we generate the data using the sample period averages of the parameter estimates we report in Table 6 assuming autocorrelation and regime persistence in the trade direction indicators. And we use the trading volume data on June 29 in simulation. To minimize simulation errors, we generate the data 50 times, and estimate the model parameters and trade direction indicators for data sample we generate and compute the simulated sample averages. We then

[14] We incorporate the square root of trading volume following Hasbrouck (2004). The estimated relation between order size and price impact is then concave.

compare the estimation results obtained from MLE methods with the Bayesian MCMC methods we propose.

**[Table 5 in here]**

Table 5 presents the results from this simulation Here $\sigma_u$ is the estimate of the standard deviation of changes in the (log) efficient price and $c$ is that of the order processing component of the (log) half spread; $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively. Finally, $P$ and $Q$ are the transition probabilities of the latent regime switching process for trade direction indicators. The "Bayesian" results are those we obtain from the Bayesian Tailored Random-walk Metropolis-Hastings Gibbs sampler in which the trade direction indicators are both conditionally simulated and autocorrelated. We generate results using 12,000 sweeps of the sampler, discarding the first 2,000. We obtain the alternative "MLE" estimates using classical MLE method to estimate the regime switching models. Once more we find that both methods provide accurate estimates of the true parameter values we use in the simulations.

### 4.6 The extended Glosten and Harris model with autocorrelated $q_t$: gold futures contract estimation

Following confirmation of the integrity of the selected estimation methodology on the basis of the simulations in section 4.5, we proceed to obtain daily estimates of the GH model with autocorrelated trade direction indicators using the gold futures data. As before, the assumption is that the parameters of this model are constant during the day may change across days. The Bayesian estimations follow an identical approach to simulation as described in section 4.2[15]. We compare the following two methods: the MLE incorporating regime switching (RS) and our extended MCMC approach.

Prior to a detailed discussion of the empirical results, we report the results of a model comparison among three models: the Roll model, the extended Roll model, and the extended GH model. Specifically, we compute:

[15] Table A.1 (the column 'GH (Bayesian)) indicates that we achieve the convergence criteria for the GH model since the effective size is greater than 1,000 and Geweke's (1992) p-values exceed 0.05 for all parameters

$$-2\ln\left(\frac{\hat{L}_R}{\hat{L}_U}\right) \sim \chi_k^2$$

where $\hat{L}_R$ and $\hat{L}_U$ are the likelihood values under the restricted and unrestricted model, so in comparing the Roll model with the extended GH model, $\hat{L}_R$ and $\hat{L}_U$ are the likelihood values under the Roll and the extended GH model, respectively. As we impose parameter restrictions on $P, Q, \lambda_0$, and $\lambda_1$ in the extended GH model to obtain the Roll model, the test statistics follows a chi-square distribution with 4 degrees of freedom. We consistently reject the Roll model and the extended Roll model in favour of the extended GH model for each day in our sample[16].

**[Table 6 in here]**

In table 6 we present the estimation results from the extended GH model, which again reveal that the MLE estimates and Bayesian estimates align closely in magnitude. The difference between the daily estimates of $c$ and $\sigma_u$ we obtain from both methods exhibits an average value close to zero, and the maximum of the absolute differential value is also minuscule ($0.12 \times 10^{-4}$). Other parameter values also exhibit very small differences,[17] leading us to conclude that both methods provide the very similar estimates. Moreover, the results reveal a consistent pattern in relation to those we obtain from the Roll and extended Roll models. The estimates of the effective trading cost parameter, $c$, with a value of $0.29 \times 10^{-4}$ are significant and comparable to the corresponding estimates from the Roll model, albeit somewhat lower than those from the extended Roll model. The daily estimates of the standard deviation of changes in the (log) efficient price, $\sigma_u$, are also significant, with an average value of $0.33 \times 10^{-4}$. The transition probability estimates again closely approximate those from the extended Roll model, again indicating the presence of moderately persistent regimes for the trade direction indicator.

[16] Table A.3 reports detailed results on the model comparison among these three models.

[17] Table A.2 provides more details on this.

Our attention is drawn to two particularly noteworthy features of the parameter estimates. First, on Brexit referendum day 23 June, the relevant estimate for $c$ is 0.84 x $10^{-4}$ , which is the largest value in the sample, and that for $\sigma_u$ is 0.78 x $10^{-4}$ , an increase of more than 250% in comparison to the previous day's. Moreover, the regime persistence coefficient, $(P+Q-1)$ is 0.5, which is significantly in excess of the average estimate, 0.4. Second, although the daily $\lambda_0$ estimates are statistically significant in only around 50% of the time, those for the slope of the price impact function arising from the effect of asymmetric information, the Kyle's lambda ($\lambda_1$) parameter are positive and always statistically significant.[18] The empirical results corroborate the implications that private, fundamental-relevant information is conveyed through trading decisions and the adverse selection costs per transaction are increasing in trade size. We further discuss the economic significance of these results in terms of the decomposition of the effective spread in section 4.8.

**[Table 7 in here]**

Prior to further analysis of the trade classification indicator, we pause to undertake some additional analysis, of the extended GH model. We motivate this on the basis of the extreme values we observe for the transition probability estimates on two dates, namely 0.9789 and 0.9278 on 19th June and 23rd June, respectively. In comparison to the remaining (sell-side) transition probability estimates we obtain, these parameter values appear as potentially implausible outliers. We conjecture that this may relate to our assumption that the model parameters remain constant throughout the day. As the above dates lie in very close proximity to the Brexit referendum, it may be implausible to impose such a restrictive assumption on the parameters. To mitigate any concerns over model misspecification, we re-estimate the extended GH model on the 19th and 23rd June using MLE methods after dividing each of these days into 10,000 trade intervals, with the resulting estimates given in Table 7. Overall, we find that the parameter estimates exhibit significant intraday variation on these two dates, with (sell-side) transition probabilities ranging from 0.6812 to 0.9756 on 19th June and 0.668 to 0.9656 on 23rd June, respectively. As such, we conclude that the initial estimates we provide in Table 6 may indeed reflect model misspecification, arising from the assumption of

[18] For an illustration see Figure 2.

constant daily parameter values.

Specifically, panel A of Table 7 reports the empirical results for 19th June. A significant change in market sentiment occurs early in the trading session on 19th June 19, after the publication of an influential survey favoring the outcome of the U.K. voting to remain in the EU. A reduction in market anxiety over a Brexit initiates an initial sell-off in safe haven assets, including gold in the early trades, and this trading behavior appears to be manifest in the estimates of two parameters in particular. First, in a much higher value for the transaction cost parameter ($c$ = 6.469), and second in that capturing the persistence of the transition probability of sell orders ($Q$ = 0.995) relative to that of buy orders ($P$ = 0.778) in the first 10,000 trades. After the first 10,000 trades, overall trading activity within subsequent 10,000 trade buckets appears relatively balanced, with similar $P$ and $Q$ parameter estimates and a much smaller $c$, albeit the $c$ estimate for trades in the 30,000 to 40,000 interval is also higher. We conclude that the extreme differences in the transition probabilities reported in Table 6 seem to mainly reflect the impact of the first 10,000 trades.

Panel B of Table 7 reports the empirical results for 23rd June, the day of the Brexit referendum. The time varying nature of the transition probability estimates over each 10,000 trade interval throughout this day clearly reveals the uncertainty relating to the outcome of the voting process. Initially, selling pressure in gold futures appears much higher with the first half of the trading day generating a very high persistence in the transition probability of sell orders. In contrast, gold futures buying pressure appears to be manifest during the second half of the trading day with the transition probability of buy orders evidencing more persistence. This structural break in the transition probability, combined with the fact that gold prices increase in uncertain times, captures the dynamics of the information flow relating to the Brexit vote result. We attribute the extreme differences in the two transition probabilities we report in table 6 to this structural break in the trading process.

To mitigate the effect of the early trading distortions on parameter estimates evident in panel A and the structural break in panel B of table 7, we decide to use the relatively moderate average estimates in the analyses of trade classification and the economic decomposition of the bid-ask spread we conduct in the next section..

### 4.7 Classification of trades for buy-sell indicator: the extended GH model

Easley, Lopez de Prado, O'Hara (2016) propose a new conceptual framework for classifying trades. They adopt the perspective of a Bayesian statistician who has priors regarding the status of the unobservable information (e.g., $q_t$), and is trying to extract investors underlying trading intentions from observable trade data. Our claim is that the empirical market microstructure models we use in this paper can provide plausible approximations to the Bayesian trade classification approach which constitutes their ideal. In particular, we maintain that we may interpret the Markov switching process in the Roll or GH models as the underlying process governing the evolution of the unobservable variables $q_t$, and the measurement equations as a plausible data generating process for the observed data relating to $q_t$. Thus, we can use estimates of the autocorrelated trade direction indicators, $q_t$, as our model-consistent trade classification algorithm.

In order to provide appropriate benchmarks with which to compare our results on the classified trades, we proceed to classify trades using the standard Tick rule[19] and generate daily correlation estimates of classified trades using the Tick rule and our model consistent rules. Specifically, on the basis of the Roll and extended GH models, using Kim (1994)'s smoothing methods and MLE parameter estimates, we calculate whether each trade is initiated by buyer ($q_t = 1$) or seller ($q_t = -1$).[20]

**[Table 8 in here]**

Table 8 presents the daily correlation estimates of trade classifications using the Tick rule, the Roll model, and the extended GH model. Two features are noteworthy. First, the daily correlation between the Tick rule and the Roll model (labelled Roll) estimates are almost always above 0.99, indicating that the Roll model essentially classifies trades on the basis of up- or down-ticks, as in the Tick rule. Second, while the daily correlation of the Tick rule

[19] The tick rule classification uses movements in trade prices to classify a trade as either a buy or a sell. Specifically, if the transaction is above (below) the previous price, then it is a buy (sell). If there is no price change, but the previous tick change was up (down), then the trade is classified as a buy (sell).

[20] As both Bayesian and MLE methods produce almost identical results, we report only the latter. The Bayesian results are available on request. Appendix D discuses Kim's (1994) smoothing algorithms.

and extended GH model (labelled GH) estimates are also high on the majority of days, they are significantly lower on both the 19th June and 23rd June, namely 0.61 and 0.45, respectively.

We conjecture a potential explanation for the second finding is as follows. Easley, Lopez de Prado, O'Hara (2016) maintain that when the underlying data is less noisy, Tick rule classifications can be superior to other rules. However, they also show that in situations where underlying data noise is substantial or order flow is imbalanced, such as when private information motivates trading, trade classifications using the Tick rule may be unreliable. In particular, the Tick rule underestimates (overestimates) the probability of buys when the direction of order flow imbalances signals positive (negative) information. As we explain in the previous section, over the days surrounding the Brexit referendum incorporate a period of great uncertainty as reflected in the results of opinion polls regarding its outcome. For example the reduction in market anxiety relating to the possibility of a leave vote on 19th June, following the release of an influential survey, initiated a major unwinding of long positions in the gold futures market in early trading, the selling pressure generating high illiquidity costs. During the remainder of this day, trades are balanced overall. This trading pattern is reflected in the daily volatility estimate, which is 0.52, much higher than its value on most other days. The day of the Brexit referendum, 23rd June also generates an abnormally high volatility estimate of 0.78. Moreover, the transition probability estimates we obtain from the extended GH model also indicate a structural break in the trading process on this day. There is initial selling pressure as the consensus in overnight opinion polls indicates a remain outcome, but the buying pressure dominates in the latter part of trading after the Brexit vote result, consistent with evidence documenting that gold prices rise during times of economic uncertainty (Erb and Harvey (2013)). The value of Kyle's lambda on June 23 is 0.324, its highest value in our sample.

In summary, the model consistent trade direction classification algorithm based on the extended GH formulation generates very similar results to the Tick rule during normal trading periods, but in periods characterised by higher uncertainty and the existence of a potentially larger price impact of trades (closely related to order imbalances), the classifications obtained from the two methods diverge significantly. As these are precisely the circumstances under which Easley, Lopez de Prado, and O'Hara (2016) argue that the Tick rule appears most

problematic in classifying trades, this suggest our proposed extended GH methods may be useful in such an environment.

Ii is important to state that we certainly do not intend to claim that our trade classification system is in any sense superior to other rules such as the Tick and BVC rules. As Easley, Lopez de Prado, and O'Hara (2016) note, perhaps each trade classification rule manifests both strengths and weakness, depending upon market conditions and the nature of the information environment. We believe that our approach to classifying trades is particularly well- suited to situations in which researchers use a variant of state space models incorporating regime switching to model trading environments. Moreover, our approach provides easy-to-implement model-consistent trade classification algorithms using both Bayesian and Classical methods. As such, we believe they provide a useful addition to the empirical microstructure tool kit.

### 4.8 Components of the effective spread: the relative contribution of order processing and adverse selection costs

How important are adverse selection costs arising from information asymmetries as constituents of the effective bid-ask spread? We address this issue by computing the relative importance in spread composition of our estimated measure of non-informational (order processing) costs and informational asymmetry components in spread composition. To facilitate the ensuing explanation, consider the log bid-ask spread implied by our model, namely:

$$p_t = m_t + cq_t, m_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - E_{t-1}q_t\right) + u_t$$
$$\Rightarrow p_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)q_t + cq_t + u_t - \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(\mu + \rho q_{t-1}\right)$$

It follows that we can express the ask ($a_t$) and bid ($b_t$) prices in our model as:

$$a_t = m_{t-1} + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right) + c + u_t - \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(\mu + \rho q_{t-1}\right)$$
$$b_t = m_{t-1} - \left(\lambda_0 + \lambda_1\sqrt{V_t}\right) - c + u_t - \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(\mu + \rho q_{t-1}\right)$$
$$a_t - b_t = 2\left(c + \lambda_0 + \lambda_1\sqrt{V_t}\right)$$

**[Table 9 in here]**

Based on this representation of the ask and bid prices, Table 9 and Figure 3 provide the daily estimates of transaction costs for a trade measured in ticks, $S_{GH,cs}$. We calculate this as $S_{GH,cs} = sp_{GH,cs} \times \bar{P}$ where $sp_{GH,cs}$ is the log spread estimate ($a_t - b_t$ in the above model) with the daily average volume ($\bar{V}_t$) (computed as $sp_{GH,cs} = 2 \times (c + \lambda_0 + \lambda_1 \sqrt{\bar{V}_t})$ ), and $\bar{P}$ is the daily average price in ticks. We also present the log spread and the daily average price in ticks in the table. For example, $S_{GH,cs}$ on 1 May is 1.1970 (i.e., $2 \times (0.4615 \times 10^{-4}) \times (1{,}296.9/0.1)$) where 1,296.9 is the mean of the daily prices and 0.1 is the tick size on that day.

To summarise our findings, the daily estimates of transaction costs for a trade are in the region of 1.2 ticks, reflecting the high liquidity of the gold futures markets. However, there are two exceptions to these estimates on 19 and 23 June. On these two days, the daily estimates of transactions costs for a trade are significantly higher, namely 6.265 (3.299) ticks on June 19 (June 23), respectively. This reflects the illiquidity arising from the enhanced uncertainty in the trading environment. Based on the parameter estimates from Table 6, in the final two columns of table 9 present the contribution of information and non-information related components to the spread (for an illustration see Figure 4). Specifically, the proportion of the spread attributable to the order processing cost component is $\text{TC} = c_0 / \left( c_0 + \left( \lambda_0 + \lambda_1 \sqrt{V_t} \right) \right)$, and the proportion arising from adverse selection costs, the information asymmetry component, is $\text{IC} = \left( \lambda_0 + \lambda_1 \sqrt{V_t} \right) / \left( c_0 + \left( \lambda_0 + \lambda_1 \sqrt{V_t} \right) \right)$. These two components are calculated by including the estimate of $\lambda_0$ only when it is statistically significant. In summary, the proportion of effective trading costs arising from non-information related components are higher than those from the information components on all days. In most cases, the former lies in the range from 55% to 70%. The exception is on June 19 where the proportion attributable to the non-information components increases to 90%. However, in general, the proportion contributed by the information related components in the gold futures market is sizeable and significant. The extended Glosten-Harris type models we estimate identify permanent price impacts arising from asymmetric information as movements in the efficient price, which is ultimately reflected in transaction prices. Viewed from this perspective, given

the nature of the gold futures market, it is perhaps not too surprising that 35 to 50% of the average bid-ask spreads reflects a compensation for bearing adverse selection risk.

## 5. CONCLUSION

Generating accurate measures of liquidity, and measuring trading costs and the price impact of trades is difficult when an absence of quotes makes classifying investor's trading intentions problematic. Existing literature incorporates several proposed resolutions to this problem but the dynamic evolution of trading mechanisms and the advent of electronic platforms creates further difficulties for some of these approaches. For example, Easley, Lopez de Prado and O'Hara (2016) maintain that in electronic limit order markets, often manifesting order cancellation rates of 98% or more, trade classification algorithms based on proximity to bid and ask quotes are severely compromised.

To overcome these data limitation, Hasbrouck (2004) proposes a new Bayesian approach by assuming an i.i.d. normal distribution for price innovations and latent independent trade indicators. Subsequently, Van der Wel, Menkveld, and Sarkar (2009) develop the equivalent classical maximum likelihood estimation (MLE) methods by mapping the Roll model onto Kim and Nelson's (1999) regime switching state space model. The first contribution of this paper is to develop easy-to-implement Bayesian and MLE estimators by extending both Hasbrouck (2004) and Van der Wel, Menkveld, and Sarkar (2009) to simultaneously accommodate several of the features which are omitted from these models, namely unbalanced and autocorrelated order flow and informational asymmetries. These omissions are evaluated in Chen, Linton, Schneeberger, and Yi (2016), but the present paper is the first to undertake a comprehensive empirical implementation which addresses these drawbacks. The second contribution of this paper is to provide robust trade direction classification mechanisms without recourse to quotes. Our proposed classification systems utilise both Bayesian MCMC methods and classical filtering and smoothing algorithms for latent trade direction indicators.

Simulation results reveal that the methods we propose are reliable. For purposes of illustration, we analyse the empirical behaviour of gold futures prices from the CME contract during a period of market uncertainty surrounding the UK's Brexit referendum in 2016. This

analysis reveals several noteworthy features. First, trade direction indicators appear highly autocorrelated, leading to measured bid-ask spreads being larger, in an economically meaningful sense, than those obtained from alternative estimates employing independent trade direction indicators. Second, they provide statistical support for asymmetric information models of the type proposed by Glosten and Harris (1988) in the presence of latent and autocorrelated trade direction indicators, thereby evidence that the trade impact coefficients implied by the asymmetric information model, which reflect Kyle's lambda, are important elements of liquidity. Third, they reveal that the trade classifications we obtain from the Roll model used in Hasbrouck (2004) and the Tick rule are essentially identical. Finally, our model consistent trade classification algorithm provides very similar results to the Tick rule during normal trading periods. However, in the presence of greater uncertainty when trading potentially generates a greater price impact (resulting from order flow imbalances), our trade classification indicator often diverges significantly from those using the Tick rule. Easley, Lopez de Prado, and O'Hara (2016) maintain that Tick rule classifications appear particularly problematic in periods of high volatility exhibiting imbalances in order flow. We believe the approach to trade classification we propose shows some promise in this type of trading environment. However, we certainly do not claim that our trade classification system is superior to other rules. As Easley, Lopez de Prado and O'Hara (2016) note, each trade classification rule may demonstrate both strengths and weakness, depending on the underlying market characteristics. Instead, we maintain that our approach may be best suited to classifying trades consistently in environments where a variant of state space models with regime switching yields a realistic approximation to the trading conditions. Moreover, our methods have the advantage of providing easy-to-implement model consistent trade classification algorithms using both Bayesian and Classical estimation methods. As such, we believe they may be a useful addition to the empirical microstructure tool kit.

## REFERENCES


Brennan, M.J., and A. Subrahmanyam, 1996, Market Microstructure and Asset Pricing: On the Compensation for Illiquidity in Stock Returns, *Journal of Financial Economics*, 41, 441-464.

Brogaard, J., Hendershott, T., and Riordan, R., 2014. High Frequency Trading and Price Discovery. *Review of Financial Studies*, 27 (8), 2267–2306,

Chib, S., and S. Ramamurthy, 2010, 'Tailored Randomized Block MCMC Methods with Application to DSGE Models, *Journal of Econometrics*, 155(1), 19-38.

Chen, X., Linton, O., Schneeberger, S., and Y. Yi, 2016, Simple Nonparametric Estimators for the Bid-Ask Spread in the Roll Model, March 30, Working paper

Choi, J.Y., Salandro, D., and K. Shastri, 1988, On the Estimation of Bid-ask Spreads: Theory and Evidence, *Journal of Financial and Quantitative Analysis*, 23, 219-30.

Easley, D, López de Prado, M., and M. O'Hara, 2012, Flow Toxicity and Liquidity in a High-frequency World, *The Review of Financial Studies*, 25(5), 1457–1493

Easley, D, Lopez de Prado, M., and M. O'Hara, 2016, Discerning Information from Trade Data, *Journal of Financial Economics*, 120(2), 269-285.

Erb, C.B., and C. R. Harvey, 2013, the Golden Dilemma, NBER working paper.

Foucault, T., Pagano, M., and A. Röell, 2013, Market Liquidity: Theory, Evidence, and Policy, Oxford University Press.

Geweke, J., 1992, Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments, Bayesian Statistics 4, Oxford University Press.

Glosten, L.R., and L.E. Harris, 1988, Estimating the Components of the Bid-ask Spread, *Journal of Financial Economics*, 21, 123–142.

Glosten, L. R., and P. Milgrom, 1985, Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders, *Journal of Financial Economics*, 14, 71-100.

Hasbrouck, J, 1991, Measuring the Information Content of Stock Trades, *Journal of Finance*, 46, 179-207.

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

Hasbrouck, J., 2007, Empirical Market Microstructure, Oxford University Press.

Hauptfleisch, M., Putniņš, T.J., and B. Lucey, 2016, Who Sets the Price of Gold? London or New York, *Journal of Futures Markets*, 36(6), 564-586.

Hendershott, T., and A. J. Menkveld, 2014, Price Pressures, *Journal of Financial Economics*, 114, 405-423.

Holden, C W., and S. Jacobsen., 2014, Liquidity Measurement Problems in Fast, Competitive Markets: Expensive and Cheap Solutions, *Journal of Finance*, 69(4), 1747-1785.

Holden, C. W., Jacobsen, S.E., and A. Subrahmanyam, 2013, The Empirical Analysis of Liquidity, Foundations and Trends in Finance, 8(4), 263-365.

Kim, C. J., 1994, Dynamic Linear Models with Markov-switching, *Journal of Econometrics*, 60,1-22.

Kim, C. J., and C. R. Nelson, 1999, State-space Models with Regime Switching: Classical and Gibbs-sampling Approaches with Applications, MIT Press.

Kim, S., Shephard, N., and S. Chib, 1998, Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65, 361–393.

Kyle, A., 1985, Continuous Auctions and Insider Trading, *Econometrica*, 53, 1315-1336.

Labuszewsk, J.W., Nyhoff, J.E., Co, R., and P. E. Peterson, 2010, The CME group risk management handbook, John Wiley & Sons.

Lee, C.M.C., and M. J. Ready, 1991, Inferring Trade Direction from Intraday Data, *Journal of Finance*, 46, 733-46.

Madhavan, A., Richardson, M., and M. Roomans, 1997, Why Do Security Prices Change?

*Review of Financial Studies*, 10, 1035-64.

Marshall, B.R., Nguyen, N.H., and N. Visaltanachoti, 2013, Liquidity Commonality in Commodities, 37 (1), *Journal of Banking and Finance*, 11-20.

Menkveld, A.J., Koopman, S.J., Lucas, A., 2007. Modelling Round-the-clock Price Discovery for Crosslisted Stocks Using State Space Methods. *Journal of Business and Economic Statistics* 25, 213–225.

Menkveld, A.J., 2013. High-frequency Trading and the New Market Makers. *Journal of Financial Markets* 16, 712–740.

Panayides M., Shohfi, T., and J. Smith, 2014, Comparing Trade Flow Classification Algorithms in the Electronic Era: The Good, the Bad, and the Uninformative, September, Working paper.

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

Sadka, R., 2006, Momentum and Post Earnings Announcement Drift Anomalies: the Role of Liquidity risk, *Journal of Financial Economics*, 80, 309-349.

Van der Wel, M., Menkveld, A. J. and A. Sarkar, 2009, Are Market Makers Uninformed and Passive? Signing Trades in the Absence of Quotes, Federal Reserve Bank of New York, Staff Reports, no.395, September.

## Table 1. Contract Descriptions and Summary Sample Statistics

| Contract | Gold futures |
|---|---|
| Expiration months | June, 2016 (GCM6) and August, 2016 (GCQ6) |
| Trading sample months | May, 2016 - June, 2016 |
| Numbers of trading days | 44 |
| Avg. price | 1269.0 |
| Price units | U.S. dollars and cents per troy ounce |
| Tick | $0.10 per troy ounce |
| Avg. tick/price | 0.008% |
| Size of contract | 100 troy ounce |
| Avg. dollar value | $126,900.00 |
| Std. Dev. Of price change (log price X 10,000) | 0.5579 |
| Std. Dev. Of price change (ticks) | 7.0849 |
| Avg. daily trades | 58,205 |
| Avg. time between trades (seconds) | 1.618808 |

## Table 2. The Roll Model: MLE and Bayesian Methods

$p_t = m_t + cq_t, m_t = m_{t-1} + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1,1\}$ where $\Pr[q_t = 1 \mid q_{t-1} = 1] = \Pr[q_t = -1 \mid q_{t-1} = -1] = 0.5$

This table provides daily parameter estimates of the above model for the gold futures contract traded on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. Electronic trading on CME Globex is available virtually 24 hours a day from Sunday 6:00 p.m. through Friday 5:00 p.m. EST, with a 60-minute break each day beginning at 5:00 p.m EST. Thus, we assume the trading activity on any given date starts at 6:00 p.m. and finishes at 5:00 p.m. EST next day. For example, the May 1 trading starts at 6:00 p.m. on May 1 and finishes at 5:00 p.m. on May 2. $\sigma_u$ is the standard deviation of the log efficient price changes, and $c$ is the (log) half spread. Estimates labelled "Roll(Bayesian)" are Hasbrouck (2004)'s Gibbs sampler estimates in which trade direction indicators are conditionally simulated. Estimates labelled " Roll(MLE)" are MLE estimates based on the Roll model. In this table, 10,000 fold of $c$ and $\sigma_u$ estimates are reported. And to guarantee a correct statistical inference, we start all Bayesian estimations with 100,000 burn-in period and 250,000 total numbers of iterations and increase these numbers by 10,000 until the convergence criteria for the inefficiency factors and the Geweke (1992)'s diagnostic tests are satisfied simultaneously for all parameters.

| | | Roll (MLE) | | Roll (Bayesian) | |
|---|---|---|---|---|---|
| Date | Parameters | $c \times 10,000$ | $\sigma_u \times 10,000$ | $c \times 10,000$ | $\sigma_u \times 10,000$ |
| 1-May | Mean | **0.2108** | **0.4510** | **0.2107** | **0.4510** |
| | Std | [0.0023] | [0.0026] | [0.0026] | [0.0023] |
| 2-May | Mean | **0.2441** | **0.4346** | **0.2442** | **0.4346** |
| | Std | [0.0024] | [0.0027] | [0.0025] | [0.0024] |
| 3-May | Mean | **0.2273** | **0.4511** | **0.2273** | **0.4511** |
| | Std | [0.0023] | [0.0025] | [0.0025] | [0.0023] |
| 4-May | Mean | **0.2237** | **0.4433** | **0.2237** | **0.4434** |
| | Std | [0.0025] | [0.0028] | [0.0027] | [0.0025] |
| 5-May | Mean | **0.2181** | **0.4674** | **0.2181** | **0.4675** |
| | Std | [0.0021] | [0.0023] | [0.0023] | [0.0021] |
| 8-May | Mean | **0.2112** | **0.4271** | **0.2113** | **0.4270** |
| | Std | [0.0024] | [0.0026] | [0.0025] | [0.0023] |
| 9-May | Mean | **0.2135** | **0.4363** | **0.2135** | **0.4363** |
| | Std | [0.0026] | [0.0028] | [0.0028] | [0.0025] |
| 10-May | Mean | **0.2421** | **0.4113** | **0.2421** | **0.4114** |
| | Std | [0.0029] | [0.0032] | [0.0031] | [0.0030] |
| 11-May | Mean | **0.2390** | **0.4274** | **0.2390** | **0.4273** |
| | Std | [0.0026] | [0.0029] | [0.0027] | [0.0026] |
| 12-May | Mean | **0.2302** | **0.4392** | **0.2302** | **0.4392** |
| | Std | [0.0026] | [0.0029] | [0.0028] | [0.0026] |
| 15-May | Mean | **0.3073** | **0.3688** | **0.3061** | **0.3702** |
| | Std | [0.0033] | [0.0043] | [0.0062] | [0.0063] |
| 16-May | Mean | **0.2505** | **0.4147** | **0.2506** | **0.4147** |
| | Std | [0.0027] | [0.0031] | [0.0029] | [0.0028] |
| 17-May | Mean | **0.2463** | **0.4616** | **0.2463** | **0.4616** |
| | Std | [0.0024] | [0.0026] | [0.0025] | [0.0024] |
| 18-May | Mean | **0.2614** | **0.4127** | **0.2616** | **0.4125** |
| | Std | [0.0029] | [0.0032] | [0.0032] | [0.0031] |
| 19-May | Mean | **0.2632** | **0.3982** | **0.2637** | **0.3977** |
| | Std | [0.0039] | [0.0044] | [0.0048] | [0.0047] |
| 22-May | Mean | **0.2403** | **0.3942** | **0.2407** | **0.3939** |
| | Std | [0.0038] | [0.0041] | [0.0041] | [0.0038] |
| 23-May | Mean | **0.3548** | **0.3062** | **0.3547** | **0.3062** |

| | | | | | |
|---|---|---|---|---|---|
| | Std | [0.0016] | [0.0024] | [0.0013] | [0.0011] |
| 24-May | Mean | **0.3463** | **0.3071** | **0.3463** | **0.3072** |
| | Std | [0.0017] | [0.0027] | [0.0015] | [0.0013] |
| 25-May | Mean | **0.2304** | **0.4424** | **0.2304** | **0.4423** |
| | Std | [0.0031] | [0.0033] | [0.0033] | [0.0030] |
| 26-May | Mean | **0.2503** | **0.4331** | **0.2503** | **0.4330** |
| | Std | [0.0032] | [0.0035] | [0.0033] | [0.0032] |
| 29-May | Mean | **0.2765** | **0.4867** | **0.2766** | **0.4867** |
| | Std | [0.0037] | [0.0043] | [0.0040] | [0.0039] |
| 30-May | Mean | **0.3359** | **0.3473** | **0.3356** | **0.3476** |
| | Std | [0.0026] | [0.0039] | [0.0026] | [0.0026] |
| 31-May | Mean | **0.3553** | **0.3256** | **0.3553** | **0.3257** |
| | Std | [0.0018] | [0.0029] | [0.0016] | [0.0015] |
| 1-Jun | Mean | **0.2734** | **0.4109** | **0.2739** | **0.4105** |
| | Std | [0.0041] | [0.0047] | [0.0052] | [0.0051] |
| 2-Jun | Mean | **0.2425** | **0.5173** | **0.2425** | **0.5173** |
| | Std | [0.0024] | [0.0026] | [0.0026] | [0.0024] |
| 5-Jun | Mean | **0.2398** | **0.4202** | **0.2398** | **0.4202** |
| | Std | [0.0031] | [0.0033] | [0.0031] | [0.0029] |
| 6-Jun | Mean | **0.3456** | **0.3020** | **0.3455** | **0.3020** |
| | Std | [0.0018] | [0.0029] | [0.0016] | [0.0014] |
| 7-Jun | Mean | **0.3400** | **0.2914** | **0.3400** | **0.2914** |
| | Std | [0.0016] | [0.0026] | [0.0014] | [0.0012] |
| 8-Jun | Mean | **0.3304** | **0.3001** | **0.3304** | **0.3002** |
| | Std | [0.0017] | [0.0028] | [0.0015] | [0.0014] |
| 9-Jun | Mean | **0.3227** | **0.3074** | **0.3226** | **0.3075** |
| | Std | [0.0019] | [0.0030] | [0.0018] | [0.0016] |
| 12-Jun | Mean | **0.2350** | **0.4005** | **0.2352** | **0.4004** |
| | Std | [0.0027] | [0.0030] | [0.0029] | [0.0027] |
| 13-Jun | Mean | **0.3273** | **0.2840** | **0.3273** | **0.2841** |
| | Std | [0.0014] | [0.0023] | [0.0012] | [0.0011] |
| 14-Jun | Mean | **0.2248** | **0.4496** | **0.2248** | **0.4496** |
| | Std | [0.0024] | [0.0027] | [0.0026] | [0.0024] |
| 15-Jun | Mean | **0.2463** | **0.4026** | **0.2464** | **0.4025** |
| | Std | [0.0022] | [0.0024] | [0.0023] | [0.0022] |
| 16-Jun | Mean | **0.2374** | **0.3886** | **0.2375** | **0.3885** |
| | Std | [0.0029] | [0.0032] | [0.0030] | [0.0029] |
| 19-Jun | Mean | **0.1846** | **0.5301** | **0.1845** | **0.5302** |
| | Std | [0.0026] | [0.0026] | [0.0031] | [0.0024] |
| 20-Jun | Mean | **0.2169** | **0.4635** | **0.2169** | **0.4635** |
| | Std | [0.0024] | [0.0026] | [0.0027] | [0.0024] |
| 21-Jun | Mean | **0.2151** | **0.4446** | **0.2151** | **0.4446** |
| | Std | [0.0028] | [0.0031] | [0.0031] | [0.0028] |
| 22-Jun | Mean | **0.2087** | **0.5066** | **0.2087** | **0.5066** |
| | Std | [0.0027] | [0.0029] | [0.0032] | [0.0027] |
| 23-Jun | Mean | **0.3313** | **0.9604** | **0.3312** | **0.9604** |
| | Std | [0.0027] | [0.0025] | [0.0030] | [0.0023] |
| 26-Jun | Mean | **0.2081** | **0.5356** | **0.2081** | **0.5356** |
| | Std | [0.0022] | [0.0023] | [0.0026] | [0.0021] |
| 27-Jun | Mean | **0.2396** | **0.4474** | **0.2396** | **0.4474** |
| | Std | [0.0025] | [0.0028] | 0.0027 | 0.0026 |
| 28-Jun | Mean | **0.2280** | **0.4369** | **0.2280** | **0.4369** |
| | Std | [0.0025] | [0.0028] | 0.0027 | 0.0025 |
| 29-Jun | Mean | **0.2231** | **0.4189** | **0.2230** | **0.4189** |
| | Std | [0.0025] | [0.0028] | 0.0026 | 0.0024 |
| Average | Mean | 0.2591 | 0.4252 | 0.2591 | 0.4251 |

## Table 3. The extended Roll Model (MS) : Simulation Study I

$$p_t = m_t + cq_t, m_t = m_{t-1} + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1,1\} \text{ where } \Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$$

This table provides parameter estimates of the above model using simulated data. In order to establish a sound basis for a plausible data generating process in the simulations, we generate the data using the averages of parameter estimates (labelled as TRUE) reported in Table 4.

$\sigma_u$ is the standard deviation of the (log) efficient price changes: $c$ is the (log) half spread. $P$ and $Q$ are transition probabilities. Estimates labeled "MS (Bayesian)" are our single-move Gibbs sampler estimates in which trade direction indicators are conditionally simulated and autocorrelated. Estimates labelled "MS (MLE)" are MLE estimates based on the extended Roll model. To minimize simulation errors, we simulate data 50 times, and estimate the parameters and trade direction indicators of these models for each generated data sample and compute simulated sample averages of $c$, $\sigma_u$, $P$ and $Q$. In this table, 10,000 fold of $c$ and $\sigma_u$ estimates are reported.

| | | TRUE | MS (MLE) | | MS (Bayesian) | |
|---|---|---|---|---|---|---|
| Model | Parameters | Average estimate | Estimate | STD | Estimate | STD |
| MS | $c \times 10,000$ | 0.43 | **0.4297** | 0.0021 | **0.4297** | 0.0019 |
| | $\sigma_u \times 10,000$ | 0.35 | **0.3498** | 0.0025 | **0.3499** | 0.0014 |
| | $P$ | 0.70 | **0.6990** | 0.0043 | **0.6991** | 0.0042 |
| | $Q$ | 0.72 | **0.7193** | 0.0041 | **0.7194** | 0.0039 |

## Table 4. The extended Roll Model (MS) : MLE and Bayesian Methods

$$p_t = m_t + cq_t, m_t = m_{t-1} + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1,1\} \text{ where } \Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$$

This table provides daily parameter estimates of the above model for the gold futures contract traded on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper) $\sigma_u$ is the standard deviation of the (log) efficient price changes: $c$ is the (log) half spread. $P$ and $Q$ are transition probabilities. Estimates labeled "MS (Bayesian)" are our single-move Gibbs sampler estimates in which trade direction indicators are conditionally simulated and autocorrelated. Estimates labelled "MS (MLE)" are MLE estimates based on the extended Roll model. In this table 10,000 fold of $c$ and $\sigma_u$ estimates are reported. And to guarantee a correct statistical inference, we start all Bayesian estimations with 100,000 burn-in period and 250,000 total numbers of iterations and increase these numbers by 10,000 until the convergence criteria for the inefficiency factors and the Geweke(1992)'s diagnostic tests are satisfied simultaneously for all parameters.

| | | MS(MLE) | | | | MS(Bayeisan) | | | |
|---|---|---|---|---|---|---|---|---|---|
| Date | Parameters | $c \times 10,000$ | $\sigma_u \times 10,000$ | $P$ | $Q$ | $c \times 10,000$ | $\sigma_u \times 10,000$ | $P$ | $Q$ |
| 1-May | Mean | **0.3823** | **0.3565** | **0.7128** | **0.7241** | **0.3823** | **0.3566** | **0.7129** | **0.7240** |
| | Std | [0.0021] | [0.0027] | [0.0058] | [0.0058] | [0.0021] | [0.0016] | [0.0058] | [0.0055] |
| 2-May | Mean | **0.3874** | **0.3481** | **0.6848** | **0.7000** | **0.3875** | **0.3481** | **0.6848** | **0.7002** |
| | Std | [0.0018] | [0.0025] | [0.0054] | [0.0052] | [0.0018] | [0.0013] | [0.0053] | [0.0050] |
| 3-May | Mean | **0.3999** | **0.3492** | **0.7106** | **0.7153** | **0.3999** | **0.3493** | **0.7103** | **0.7158** |
| | Std | [0.0018] | [0.0025] | [0.0054] | [0.0054] | [0.0018] | [0.0013] | [0.0053] | [0.0052] |
| 4-May | Mean | **0.3900** | **0.3463** | **0.7127** | **0.7102** | **0.3900** | **0.3464** | **0.7129** | **0.7102** |
| | Std | [0.0020] | [0.0027] | [0.0053] | [0.0054] | [0.0020] | [0.0014] | [0.0052] | [0.0053] |
| 5-May | Mean | **0.4024** | **0.3631** | **0.7139** | **0.7271** | **0.4024** | **0.3632** | **0.7136** | **0.7275** |
| | Std | [0.0018] | [0.0025] | [0.0055] | [0.0055] | [0.0018] | [0.0014] | [0.0056] | [0.0052] |
| 8-May | Mean | **0.3886** | **0.3241** | **0.7057** | **0.7441** | **0.3887** | **0.3241** | **0.7058** | **0.7440** |
| | Std | [0.0018] | [0.0025] | [0.0050] | [0.0045] | [0.0018] | [0.0011] | [0.0051] | [0.0043] |
| 9-May | Mean | **0.3856** | **0.3390** | **0.7247** | **0.7183** | **0.3856** | **0.3390** | **0.7246** | **0.7185** |
| | Std | [0.0021] | [0.0028] | [0.0055] | [0.0056] | [0.0021] | [0.0014] | [0.0054] | [0.0054] |
| 10-May | Mean | **0.3850** | **0.3266** | **0.7052** | **0.7006** | **0.3850** | **0.3266** | **0.7053** | **0.7007** |
| | Std | [0.0019] | [0.0026] | [0.0052] | [0.0053] | [0.0019] | [0.0012] | [0.0052] | [0.0053] |
| 11-May | Mean | **0.3883** | **0.3365** | **0.6897** | **0.7090** | **0.3883** | **0.3365** | **0.6901** | **0.7088** |
| | Std | [0.0018] | [0.0026] | [0.0052] | [0.0050] | [0.0018] | [0.0013] | [0.0052] | [0.0049] |
| 12-May | Mean | **0.3891** | **0.3449** | **0.6928** | **0.7181** | **0.3891** | **0.3450** | **0.6929** | **0.7180** |
| | Std | [0.0020] | [0.0027] | [0.0055] | [0.0051] | [0.0020] | [0.0014] | [0.0055] | [0.0049] |
| 15-May | Mean | **0.3909** | **0.3293** | **0.6695** | **0.6860** | **0.3910** | **0.3293** | **0.6696** | **0.6861** |
| | Std | [0.0017] | [0.0025] | [0.0052] | [0.0051] | [0.0017] | [0.0012] | [0.0053] | [0.0050] |
| 16-May | Mean | **0.3850** | **0.3332** | **0.6909** | **0.6936** | **0.3850** | **0.3332** | **0.6909** | **0.6937** |
| | Std | [0.0018] | [0.0025] | [0.0052] | [0.0051] | [0.0018] | [0.0012] | [0.0051] | [0.0051] |
| 17-May | Mean | **0.4002** | **0.3686** | **0.6885** | **0.6976** | **0.4002** | **0.3686** | **0.6885** | **0.6977** |
| | Std | [0.0019] | [0.0025] | [0.0056] | [0.0055] | [0.0019] | [0.0015] | [0.0056] | [0.0054] |
| 18-May | Mean | **0.3996** | **0.3278** | **0.6846** | **0.7088** | **0.3997** | **0.3279** | **0.6847** | **0.7089** |
| | Std | [0.0016] | [0.0022] | [0.0045] | [0.0042] | [0.0016] | [0.0011] | [0.0045] | [0.0041] |
| 19-May | Mean | **0.3907** | **0.3238** | **0.6822** | **0.7107** | **0.3907** | **0.3239** | **0.6826** | **0.7106** |
| | Std | [0.0021] | [0.0028] | [0.0057] | [0.0053] | [0.0020] | [0.0013] | [0.0057] | [0.0052] |
| 22-May | Mean | **0.3897** | **0.3075** | **0.7042** | **0.7312** | **0.3897** | **0.3075** | **0.7045** | **0.7309** |
| | Std | [0.0019] | [0.0025] | [0.0054] | [0.0050] | [0.0019] | [0.0011] | [0.0054] | [0.0049] |
| 23-May | Mean | **0.4098** | **0.3035** | **0.7051** | **0.7187** | **0.4098** | **0.3034** | **0.7054** | **0.7186** |
| | Std | [0.0015] | [0.0021] | [0.0048] | [0.0046] | [0.0016] | [0.0010] | [0.0047] | [0.0045] |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 24-May | Mean | **0.4025** | **0.3055** | **0.7208** | **0.7053** | **0.4025** | **0.3055** | **0.7207** | **0.7054** |
| | Std | [0.0018] | [0.0023] | [0.0049] | [0.0052] | [0.0018] | [0.0010] | [0.0049] | [0.0052] |
| 25-May | Mean | **0.4055** | **0.3369** | **0.7081** | **0.7255** | **0.4055** | **0.3369** | **0.7080** | **0.7257** |
| | Std | [0.0021] | [0.0029] | [0.0056] | [0.0053] | [0.0021] | [0.0014] | [0.0055] | [0.0052] |
| 26-May | Mean | **0.4085** | **0.3356** | **0.7040** | **0.7075** | **0.4085** | **0.3357** | **0.7042** | **0.7074** |
| | Std | [0.0020] | [0.0027] | [0.0052] | [0.0051] | [0.0020] | [0.0013] | [0.0052] | [0.0051] |
| 29-May | Mean | **0.4116** | **0.4079** | **0.6459** | **0.6978** | **0.4117** | **0.4081** | **0.6464** | **0.6979** |
| | Std | [0.0034] | [0.0041] | [0.0098] | [0.0087] | [0.0034] | [0.0027] | [0.0096] | [0.0079] |
| 30-May | Mean | **0.4023** | **0.3312** | **0.6935** | **0.6882** | **0.4023** | **0.3313** | **0.6937** | **0.6881** |
| | Std | [0.0020] | [0.0027] | [0.0052] | [0.0053] | [0.0019] | [0.0013] | [0.0052] | [0.0053] |
| 31-May | Mean | **0.4080** | **0.3208** | **0.6865** | **0.6933** | **0.4080** | **0.3209** | **0.6866** | **0.6933** |
| | Std | [0.0018] | [0.0025] | [0.0050] | [0.0049] | [0.0018] | [0.0011] | [0.0049] | [0.0048] |
| 1-Jun | Mean | **0.4039** | **0.3362** | **0.6967** | **0.6979** | **0.4040** | **0.3362** | **0.6969** | **0.6979** |
| | Std | [0.0023] | [0.0030] | [0.0059] | [0.0058] | [0.0022] | [0.0014] | [0.0058] | [0.0057] |
| 2-Jun | Mean | **0.7386** | **0.5375** | **0.4476** | **0.9773** | **0.7406** | **0.5375** | **0.4550** | **0.9774** |
| | Std | [0.0150] | [0.0028] | [0.0206] | [0.0013] | [0.0110] | [0.0020] | [0.0175] | [0.0011] |
| 5-Jun | Mean | **0.4045** | **0.3204** | **0.7063** | **0.7267** | **0.4045** | **0.3204** | **0.7063** | **0.7267** |
| | Std | [0.0018] | [0.0025] | [0.0052] | [0.0049] | [0.0018] | [0.0012] | [0.0052] | [0.0049] |
| 6-Jun | Mean | **0.3913** | **0.3019** | **0.6896** | **0.6881** | **0.3913** | **0.3020** | **0.6897** | **0.6882** |
| | Std | [0.0020] | [0.0027] | [0.0057] | [0.0057] | [0.0020] | [0.0012] | [0.0056] | [0.0056] |
| 7-Jun | Mean | **0.3908** | **0.2921** | **0.7036** | **0.7176** | **0.3908** | **0.2921** | **0.7039** | **0.7174** |
| | Std | [0.0017] | [0.0023] | [0.0049] | [0.0047] | [0.0017] | [0.0010] | [0.0049] | [0.0046] |
| 8-Jun | Mean | **0.3802** | **0.2984** | **0.6956** | **0.6994** | **0.3803** | **0.2984** | **0.6957** | **0.6995** |
| | Std | [0.0019] | [0.0025] | [0.0051] | [0.0051] | [0.0018] | [0.0011] | [0.0050] | [0.0049] |
| 9-Jun | Mean | **0.3795** | **0.3019** | **0.7091** | **0.6943** | **0.3795** | **0.3019** | **0.7092** | **0.6944** |
| | Std | [0.0019] | [0.0025] | [0.0052] | [0.0054] | [0.0018] | [0.0011] | [0.0051] | [0.0054] |
| 12-Jun | Mean | **0.3822** | **0.3102** | **0.7173** | **0.6964** | **0.3822** | **0.3103** | **0.7174** | **0.6963** |
| | Std | [0.0016] | [0.0023] | [0.0046] | [0.0049] | [0.0016] | [0.0010] | [0.0045] | [0.0048] |
| 13-Jun | Mean | **0.3766** | **0.2848** | **0.7065** | **0.7152** | **0.3766** | **0.2848** | **0.7068** | **0.7150** |
| | Std | [0.0016] | [0.0021] | [0.0046] | [0.0045] | [0.0016] | [0.0009] | [0.0045] | [0.0045] |
| 14-Jun | Mean | **0.4022** | **0.3423** | **0.7234** | **0.7054** | **0.4022** | **0.3423** | **0.7232** | **0.7057** |
| | Std | [0.0018] | [0.0026] | [0.0059] | [0.0061] | [0.0018] | [0.0013] | [0.0059] | [0.0063] |
| 15-Jun | Mean | **0.3808** | **0.3154** | **0.6889** | **0.6915** | **0.3809** | **0.3154** | **0.6890** | **0.6915** |
| | Std | [0.0012] | [0.0019] | [0.0037] | [0.0037] | [0.0012] | [0.0008] | [0.0036] | [0.0036] |
| 16-Jun | Mean | **0.3767** | **0.3057** | **0.7022** | **0.7132** | **0.3767** | **0.3057** | **0.7026** | **0.7129** |
| | Std | [0.0017] | [0.0023] | [0.0048] | [0.0046] | [0.0016] | [0.0010] | [0.0048] | [0.0046] |
| 19-Jun | Mean | **0.8509** | **0.5325** | **0.6238** | **0.9883** | **0.8516** | **0.5325** | **0.6244** | **0.9883** |
| | Std | [0.0146] | [0.0024] | [0.0238] | [0.0007] | [0.0133] | [0.0018] | [0.0189] | [0.0006] |
| 20-Jun | Mean | **0.3881** | **0.3629** | **0.7117** | **0.7022** | **0.3881** | **0.3630** | **0.7121** | **0.7020** |
| | Std | [0.0020] | [0.0027] | [0.0058] | [0.0060] | [0.0021] | [0.0016] | [0.0056] | [0.0059] |
| 21-Jun | Mean | **0.3830** | **0.3463** | **0.7262** | **0.6948** | **0.3830** | **0.3464** | **0.7265** | **0.6945** |
| | Std | [0.0023] | [0.0031] | [0.0060] | [0.0065] | [0.0023] | [0.0016] | [0.0056] | [0.0065] |
| 22-Jun | Mean | **0.3988** | **0.4140** | **0.7303** | **0.7289** | **0.3987** | **0.4143** | **0.7315** | **0.7280** |
| | Std | [0.0029] | [0.0033] | [0.0098] | [0.0097] | [0.0030] | [0.0025] | [0.0096] | [0.0095] |
| 23-Jun | Mean | **1.1238** | **0.7851** | **0.5974** | **0.9404** | **1.1237** | **0.7851** | **0.5981** | **0.9404** |
| | Std | [0.0047] | [0.0022] | [0.0060] | [0.0009] | [0.0043] | [0.0017] | [0.0057] | [0.0008] |
| 26-Jun | Mean | **0.6137** | **0.5204** | **0.9633** | **0.5514** | **0.6144** | **0.5205** | **0.9634** | **0.5527** |
| | Std | [0.0106] | [0.0028] | [0.0018] | [0.0174] | [0.0077] | [0.0019] | [0.0015] | [0.0164] |
| 27-Jun | Mean | **0.3827** | **0.3605** | **0.6557** | **0.7129** | **0.3827** | **0.3606** | **0.6559** | **0.7129** |
| | Std | [0.0020] | [0.0029] | [0.0070] | [0.0063] | [0.0021] | [0.0016] | [0.0069] | [0.0057] |
| 28-Jun | Mean | **0.3777** | **0.3472** | **0.7101** | **0.6823** | **0.3777** | **0.3473** | **0.7103** | **0.6822** |
| | Std | [0.0021] | [0.0028] | [0.0062] | [0.0066] | [0.0021] | [0.0015] | [0.0059] | [0.0065] |
| 29-Jun | Mean | **0.3803** | **0.3251** | **0.7091** | **0.7118** | **0.3803** | **0.3252** | **0.7094** | **0.7115** |
| | Std | [0.0018] | [0.0026] | [0.0052] | [0.0053] | [0.0018] | [0.0012] | [0.0052] | [0.0051] |
| Average | Mean | 0.4320 | 0.3570 | 0.6966 | 0.7220 | 0.4392 | 0.3651 | 0.7011 | 0.7215 |

## Table 5. The extended GH Model (GH) : Simulation Study II

$$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t \left(q_t - E\left[q_t \mid \Omega_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1,1\}$$

where $\Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P$, $\Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\lambda_t = \lambda_0 + \lambda_1 \sqrt{V_t}$

This table provides parameter estimates of the above model using simulated data. In order to establish a sound basis for a plausible data generating process in the simulations, we generate the data using the averages of parameter estimates (labelled as TRUE) reported in Table 6.

$\sigma_u$ is the standard deviation of the (log) efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively. Estimates labeled "GH (Bayesian)" are our tailored Random-walk MH sampler estimates in which trade direction indicators are conditionally simulated and autocorrelated. Estimates labelled "GH (MLE)" are MLE estimates based on the extended GH model. To minimize simulation errors, we simulate data 50 times, and estimate the parameters and trade direction indicators of these models for each generated data sample and compute simulated sample averages of $\sigma_u, c, \lambda_0, \lambda_1, P$ and $Q$. In this table, 10,000 fold of $c, \lambda_0,$ $\lambda_1$ and $\sigma_u$ estimates are reported.

| | | TRUE | GH (MLE) | | GH (Bayesian) | |
|---|---|---|---|---|---|---|
| Model | Parameters | Average estimate | Estimate | STD | Estimate | STD |
| GH | $c \times 10,000$ | 0.30 | **0.3005** | 0.0032 | **0.3005** | 0.0033 |
| | $\sigma_u \times 10,000$ | 0.33 | **0.3299** | 0.0024 | **0.3299** | 0.0013 |
| | $\lambda_0 \times 10,000$ | 0.02 | **0.0191** | 0.0047 | **0.0191** | 0.0052 |
| | $\lambda_1 \times 10,000$ | 0.10 | **0.1000** | 0.0023 | **0.1000** | 0.0021 |
| | $P$ | 0.70 | **0.6995** | 0.0035 | **0.6996** | 0.0034 |
| | $Q$ | 0.71 | **0.7098** | 0.0034 | **0.7099** | 0.0033 |

## Table 6. The extended GH Model (GH) : Estimation using MLE and Bayesian Methods

This table provides daily parameter estimates of the above model for the gold futures contract traded on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper.) $\sigma_u$ is the standard deviation of the (log) efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively, due to adverse selection. Estimates labeled "GH (Bayesian)" are our tailored random-walk MH sampler estimates in which trade direction indicators are conditionally simulated and autocorrelated. Estimates labelled "GH (MLE)" are MLE estimates based on the extended GH model. In this table, 10,000 fold of $c$, $\lambda_0$, $\lambda_1$, and $\sigma_u$ estimates are reported. And to guarantee a correct statistical inference, we start all Bayesian estimations with 100,000 burn-in period and 250,000 total numbers of iterations and increase these numbers by 10,000 until the convergence criteria for the inefficiency factors and the Geweke (1992)'s diagnostic tests are satisfied simultaneously for all parameters.

| | | GH(MLE) | | | | | | GH(Bayeisan) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Parameters | $c_{\times 10,000}$ | $\sigma_{u\times 10,000}$ | $\lambda_{0\times 10,000}$ | $\lambda_{1\times 10,000}$ | $P$ | $Q$ | $c_{\times 10,000}$ | $\sigma_{u\times 10,000}$ | $\lambda_{0\times 10,000}$ | $\lambda_{1\times 10,000}$ | $P$ | $Q$ |
| 1-May | Mean | **0.2425** | **0.3265** | **0.0409** | **0.1052** | **0.7013** | **0.7082** | **0.2426** | **0.3266** | **0.0405** | **0.1052** | **0.7017** | **0.7078** |
| | Std | [0.0030] | [0.0028] | [0.0049] | [0.0025] | [0.0036] | [0.0035] | [0.0033] | [0.0014] | [0.0064] | [0.0025] | [0.0037] | [0.0037] |
| 2-May | Mean | **0.2677** | **0.3197** | **0.0102** | **0.1113** | **0.6764** | **0.6865** | **0.2677** | **0.3198** | 0.0102 | **0.1113** | **0.6765** | **0.6864** |
| | Std | [0.0026] | [0.0026] | [0.0043] | [0.0021] | [0.0035] | [0.0034] | [0.0029] | [0.0013] | [0.0057] | [0.0022] | [0.0036] | [0.0034] |
| 3-May | Mean | **0.2719** | **0.3235** | **0.0289** | **0.1010** | **0.7022** | **0.6970** | **0.2719** | **0.3235** | **0.0290** | **0.1009** | **0.7021** | **0.6973** |
| | Std | [0.0027] | [0.0026] | [0.0043] | [0.0022] | [0.0034] | [0.0034] | [0.0030] | [0.0012] | [0.0056] | [0.0022] | [0.0035] | [0.0035] |
| 4-May | Mean | **0.2597** | **0.3206** | **0.0437** | **0.0956** | **0.7013** | **0.7044** | **0.2598** | **0.3207** | **0.0435** | **0.0956** | **0.7013** | **0.7045** |
| | Std | [0.0030] | [0.0028] | [0.0047] | [0.0022] | [0.0036] | [0.0036] | [0.0033] | [0.0013] | [0.0062] | [0.0023] | [0.0037] | [0.0037] |
| 5-May | Mean | **0.2708** | **0.3354** | **0.0207** | **0.1073** | **0.6915** | **0.7088** | **0.2708** | **0.3355** | **0.0204** | **0.1075** | **0.6920** | **0.7084** |
| | Std | [0.0027] | [0.0025] | [0.0043] | [0.0022] | [0.0034] | [0.0032] | [0.0029] | [0.0013] | [0.0055] | [0.0022] | [0.0034] | [0.0033] |
| 8-May | Mean | **0.2611** | **0.3026** | **0.0425** | **0.0864** | **0.7125** | **0.7197** | **0.2611** | **0.3026** | **0.0426** | **0.0864** | **0.7124** | **0.7199** |
| | Std | [0.0029] | [0.0025] | [0.0042] | [0.0019] | [0.0034] | [0.0032] | [0.0031] | [0.0011] | [0.0057] | [0.0019] | [0.0036] | [0.0034] |
| 9-May | Mean | **0.2444** | **0.3120** | **0.0537** | **0.0993** | **0.7150** | **0.7140** | **0.2443** | **0.3120** | **0.0538** | **0.0993** | **0.7151** | **0.7142** |
| | Std | [0.0032] | [0.0028] | [0.0050] | [0.0023] | [0.0036] | [0.0036] | [0.0035] | [0.0013] | [0.0066] | [0.0024] | [0.0037] | [0.0038] |
| 10-May | Mean | **0.2649** | **0.3016** | **0.0215** | **0.1018** | **0.6928** | **0.6947** | **0.2649** | **0.3017** | **0.0216** | **0.1018** | **0.6929** | **0.6950** |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Std | [0.0028] | [0.0026] | [0.0045] | [0.0022] | [0.0036] | [0.0036] | [0.0030] | [0.0012] | [0.0060] | [0.0022] | [0.0037] | [0.0037] |
| 11-May | Mean | **0.2723** | **0.3120** | **0.0152** | **0.1029** | **0.6838** | **0.6937** | **0.2722** | **0.3120** | **0.0151** | **0.1031** | **0.6841** | **0.6937** |
| | Std | [0.0027] | [0.0026] | [0.0044] | [0.0022] | [0.0036] | [0.0034] | [0.0029] | [0.0012] | [0.0059] | [0.0022] | [0.0037] | [0.0035] |
| 12-May | Mean | **0.2584** | **0.3166** | **0.0388** | **0.1034** | **0.6916** | **0.7048** | **0.2584** | **0.3166** | **0.0387** | **0.1035** | **0.6917** | **0.7047** |
| | Std | [0.0029] | [0.0028] | [0.0047] | [0.0023] | [0.0036] | [0.0034] | [0.0032] | [0.0013] | [0.0061] | [0.0023] | [0.0037] | [0.0035] |
| 15-May | Mean | **0.2893** | **0.3077** | **0.0291** | **0.0847** | **0.6657** | **0.6884** | **0.2893** | **0.3077** | **0.0291** | **0.0847** | **0.6660** | **0.6882** |
| | Std | [0.0026] | [0.0027] | [0.0041] | [0.0020] | [0.0037] | [0.0035] | [0.0028] | [0.0012] | [0.0053] | [0.0019] | [0.0037] | [0.0036] |
| 16-May | Mean | **0.2692** | **0.3086** | **0.0325** | **0.0949** | **0.6897** | **0.6893** | **0.2691** | **0.3086** | **0.0327** | **0.0948** | **0.6899** | **0.6894** |
| | Std | [0.0027] | [0.0027] | [0.0044] | [0.0022] | [0.0035] | [0.0035] | [0.0029] | [0.0012] | [0.0057] | [0.0022] | [0.0036] | [0.0036] |
| 17-May | Mean | **0.2785** | **0.3375** | 0.0023 | **0.1194** | **0.6772** | **0.6788** | **0.2784** | **0.3375** | 0.0026 | **0.1193** | **0.6773** | **0.6790** |
| | Std | [0.0026] | [0.0026] | [0.0044] | [0.0023] | [0.0034] | [0.0033] | [0.0028] | [0.0013] | [0.0057] | [0.0023] | [0.0034] | [0.0034] |
| 18-May | Mean | **0.2839** | **0.3057** | **0.0457** | **0.0810** | **0.6970** | **0.7007** | **0.2859** | **0.3080** | **0.0634** | **0.0680** | **0.6904** | **0.7102** |
| | Std | [0.0025] | [0.0023] | [0.0037] | [0.0016] | [0.0030] | [0.0030] | [0.0052] | [0.0016] | [0.0157] | [0.0093] | [0.0049] | [0.0048] |
| 19-May | Mean | **0.2779** | **0.3022** | **0.0391** | **0.0869** | **0.6878** | **0.7030** | **0.2780** | **0.3023** | **0.0389** | **0.0869** | **0.6877** | **0.7031** |
| | Std | [0.0031] | [0.0029] | [0.0048] | [0.0023] | [0.0040] | [0.0038] | [0.0034] | [0.0013] | [0.0065] | [0.0023] | [0.0042] | [0.0039] |
| 22-May | Mean | **0.2783** | **0.2909** | **0.0308** | **0.0851** | **0.7086** | **0.7169** | **0.2783** | **0.2909** | **0.0309** | **0.0851** | **0.7087** | **0.7169** |
| | Std | [0.0031] | [0.0027] | [0.0047] | [0.0023] | [0.0039] | [0.0037] | [0.0034] | [0.0011] | [0.0065] | [0.0023] | [0.0041] | [0.0039] |
| 23-May | Mean | **0.3097** | **0.2871** | **0.0163** | **0.0783** | **0.7040** | **0.7026** | **0.3096** | **0.2871** | **0.0164** | **0.0784** | **0.7037** | **0.7029** |
| | Std | [0.0025] | [0.0022] | [0.0037] | [0.0016] | [0.0033] | [0.0034] | [0.0027] | [0.0009] | [0.0049] | [0.0016] | [0.0034] | [0.0034] |
| 24-May | Mean | **0.2995** | **0.2877** | 0.0003 | **0.0902** | **0.7073** | **0.6925** | **0.2995** | **0.2878** | 0.0000 | **0.0903** | **0.7073** | **0.6924** |
| | Std | [0.0027] | [0.0023] | [0.0043] | [0.0020] | [0.0037] | [0.0039] | [0.0030] | [0.0010] | [0.0059] | [0.0020] | [0.0039] | [0.0041] |
| 25-May | Mean | **0.2783** | **0.3150** | **0.0267** | **0.1026** | **0.6991** | **0.7116** | **0.2783** | **0.3150** | **0.0268** | **0.1026** | **0.6992** | **0.7118** |
| | Std | [0.0034] | [0.0030] | [0.0056] | [0.0030] | [0.0041] | [0.0038] | [0.0037] | [0.0014] | [0.0074] | [0.0029] | [0.0043] | [0.0040] |
| 26-May | Mean | **0.2840** | **0.3109** | **0.0161** | **0.1012** | **0.6936** | **0.6912** | **0.2840** | **0.3109** | **0.0162** | **0.1012** | **0.6936** | **0.6914** |
| | Std | [0.0029] | [0.0027] | [0.0047] | [0.0022] | [0.0037] | [0.0038] | [0.0032] | [0.0012] | [0.0062] | [0.0022] | [0.0039] | [0.0039] |
| 29-May | Mean | **0.2783** | **0.3706** | **0.0279** | **0.1219** | **0.6524** | **0.6839** | **0.2783** | **0.3708** | **0.0276** | **0.1220** | **0.6526** | **0.6839** |
| | Std | [0.0043] | [0.0042] | [0.0074] | [0.0038] | [0.0058] | [0.0050] | [0.0048] | [0.0025] | [0.0097] | [0.0039] | [0.0059] | [0.0051] |
| 30-May | Mean | **0.2883** | **0.3094** | **0.0395** | **0.0871** | **0.6899** | **0.6906** | **0.2883** | **0.3094** | **0.0395** | **0.0871** | **0.6900** | **0.6907** |
| | Std | [0.0030] | [0.0028] | [0.0047] | [0.0022] | [0.0038] | [0.0038] | [0.0033] | [0.0012] | [0.0063] | [0.0022] | [0.0039] | [0.0040] |
| 31-May | Mean | **0.3055** | **0.3030** | **0.0319** | **0.0801** | **0.6902** | **0.6880** | **0.3056** | **0.3030** | **0.0319** | **0.0800** | **0.6903** | **0.6881** |
| | Std | [0.0029] | [0.0026] | [0.0043] | [0.0020] | [0.0037] | [0.0037] | [0.0031] | [0.0012] | [0.0058] | [0.0020] | [0.0038] | [0.0039] |
| 1-Jun | Mean | **0.2841** | **0.3155** | **0.0597** | **0.0830** | **0.7038** | **0.7000** | **0.2841** | **0.3156** | **0.0597** | **0.0829** | **0.7039** | **0.7001** |
| | Std | [0.0036] | [0.0033] | [0.0055] | [0.0029] | [0.0041] | [0.0042] | [0.0038] | [0.0014] | [0.0073] | [0.0028] | [0.0042] | [0.0044] |
| 2-Jun | Mean | **0.3428** | **0.3876** | **-0.0602** | **0.1151** | **0.6933** | **0.6910** | **0.3156** | **0.3328** | **0.0269** | **0.0850** | **0.6825** | **0.6797** |
| | Std | [0.0034] | [0.0030] | [0.0051] | [0.0023] | [0.0049] | [0.0050] | [0.0028] | [0.0012] | [0.0051] | [0.0017] | [0.0033] | [0.0034] |
| 5-Jun | Mean | **0.2854** | **0.3014** | **0.0411** | **0.0809** | **0.7075** | **0.7157** | **0.2852** | **0.3014** | **0.0415** | **0.0809** | **0.7080** | **0.7155** |
| | Std | [0.0031] | [0.0026] | [0.0044] | [0.0019] | [0.0037] | [0.0035] | [0.0033] | [0.0011] | [0.0058] | [0.0019] | [0.0038] | [0.0036] |
| 6-Jun | Mean | **0.2984** | **0.2866** | **0.0212** | **0.0770** | **0.6939** | **0.6861** | **0.2984** | **0.2867** | **0.0213** | **0.0770** | **0.6941** | **0.6862** |
| | Std | [0.0031] | [0.0028] | [0.0048] | [0.0023] | [0.0042] | [0.0044] | [0.0034] | [0.0012] | [0.0068] | [0.0023] | [0.0044] | [0.0047] |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7-Jun | Mean | **0.2918** | **0.2764** | **0.0335** | **0.0662** | **0.6963** | **0.7154** | **0.2917** | **0.2764** | **0.0336** | **0.0662** | **0.6963** | **0.7157** |
| | Std | [0.0028] | [0.0024] | [0.0039] | [0.0015] | [0.0038] | [0.0036] | [0.0030] | [0.0010] | [0.0055] | [0.0015] | [0.0040] | [0.0038] |
| 8-Jun | Mean | **0.2780** | **0.2809** | **0.0372** | **0.0744** | **0.6977** | **0.7013** | **0.2780** | **0.2809** | **0.0372** | **0.0744** | **0.6978** | **0.7014** |
| | Std | [0.0028] | [0.0025] | [0.0042] | [0.0018] | [0.0038] | [0.0037] | [0.0031] | [0.0011] | [0.0057] | [0.0018] | [0.0039] | [0.0039] |
| 9-Jun | Mean | **0.2740** | **0.2848** | **0.0473** | **0.0710** | **0.7067** | **0.7022** | **0.2740** | **0.2848** | **0.0470** | **0.0711** | **0.7067** | **0.7022** |
| | Std | [0.0029] | [0.0027] | [0.0044] | [0.0020] | [0.0037] | [0.0039] | [0.0032] | [0.0011] | [0.0060] | [0.0020] | [0.0040] | [0.0041] |
| 12-Jun | Mean | **0.2704** | **0.2900** | **0.0280** | **0.0819** | **0.6982** | **0.7003** | **0.2703** | **0.2901** | **0.0282** | **0.0818** | **0.6985** | **0.7002** |
| | Std | [0.0025] | [0.0024] | [0.0038] | [0.0016] | [0.0034] | [0.0034] | [0.0027] | [0.0010] | [0.0051] | [0.0016] | [0.0035] | [0.0037] |
| 13-Jun | Mean | **0.2808** | **0.2692** | **0.0216** | **0.0722** | **0.6994** | **0.7052** | **0.2808** | **0.2692** | **0.0216** | **0.0722** | **0.6995** | **0.7052** |
| | Std | [0.0024] | [0.0022] | [0.0036] | [0.0016] | [0.0036] | [0.0035] | [0.0026] | [0.0009] | [0.0050] | [0.0016] | [0.0037] | [0.0037] |
| 14-Jun | Mean | **0.2834** | **0.3213** | **0.0471** | **0.0792** | **0.7045** | **0.7086** | **0.2835** | **0.3215** | **0.0467** | **0.0792** | **0.7045** | **0.7086** |
| | Std | [0.0030] | [0.0028] | [0.0043] | [0.0020] | [0.0037] | [0.0037] | [0.0032] | [0.0013] | [0.0055] | [0.0019] | [0.0037] | [0.0037] |
| 15-Jun | Mean | **0.2804** | **0.2945** | -0.0012 | **0.0942** | **0.6768** | **0.6736** | **0.2805** | **0.2945** | -0.0013 | **0.0942** | **0.6770** | **0.6735** |
| | Std | [0.0018] | [0.0019] | [0.0030] | [0.0015] | [0.0027] | [0.0028] | [0.0020] | [0.0008] | [0.0039] | [0.0015] | [0.0028] | [0.0029] |
| 16-Jun | Mean | **0.2665** | **0.2867** | **0.0361** | **0.0794** | **0.6972** | **0.7111** | **0.2664** | **0.2867** | **0.0362** | **0.0794** | **0.6976** | **0.7109** |
| | Std | [0.0025] | [0.0023] | [0.0038] | [0.0017] | [0.0034] | [0.0033] | [0.0027] | [0.0010] | [0.0051] | [0.0017] | [0.0035] | [0.0034] |
| 19-Jun | Mean | **0.5552** | **0.5191** | 0.0163 | **0.0829** | **0.5173** | **0.9789** | **0.5523** | **0.5189** | 0.0170 | **0.0837** | **0.5243** | **0.9788** |
| | Std | [0.0985] | [0.0158] | [0.1181] | [0.0406] | [0.0398] | [0.0113] | [0.0201] | [0.0024] | [0.0265] | [0.0073] | [0.0199] | [0.0017] |
| 20-Jun | Mean | **0.2551** | **0.3324** | **0.0216** | **0.1089** | **0.6974** | **0.6866** | **0.2550** | **0.3325** | **0.0213** | **0.1092** | **0.6973** | **0.6867** |
| | Std | [0.0029] | [0.0027] | [0.0047] | [0.0023] | [0.0036] | [0.0038] | [0.0031] | [0.0014] | [0.0061] | [0.0023] | [0.0037] | [0.0039] |
| 21-Jun | Mean | **0.2455** | **0.3182** | **0.0431** | **0.1031** | **0.7172** | **0.6920** | **0.2454** | **0.3182** | **0.0431** | **0.1031** | **0.7171** | **0.6922** |
| | Std | [0.0034] | [0.0032] | [0.0055] | [0.0027] | [0.0039] | [0.0043] | [0.0037] | [0.0015] | [0.0070] | [0.0027] | [0.0040] | [0.0044] |
| 22-Jun | Mean | **0.2467** | **0.3794** | -0.0091 | **0.1424** | **0.6907** | **0.6939** | **0.2465** | **0.3797** | -0.0094 | **0.1424** | **0.6906** | **0.6941** |
| | Std | [0.0037] | [0.0032] | [0.0061] | [0.0032] | [0.0050] | [0.0048] | [0.0042] | [0.0025] | [0.0084] | [0.0034] | [0.0051] | [0.0049] |
| 23-Jun | Mean | **0.8224** | **0.7571** | -0.0164 | **0.2003** | **0.5713** | **0.9278** | **0.8216** | **0.7572** | -0.0166 | **0.2008** | **0.5712** | **0.9277** |
| | Std | [0.0072] | [0.0022] | [0.0095] | [0.0037] | [0.0050] | [0.0010] | [0.0068] | [0.0017] | [0.0095] | [0.0037] | [0.0048] | [0.0009] |
| 26-Jun | Mean | **0.2759** | **0.4503** | **-0.1332** | **0.1794** | **0.7824** | **0.6354** | **0.2600** | **0.3801** | **-0.0108** | **0.1281** | **0.6979** | **0.6735** |
| | Std | [0.0108] | [0.0034] | [0.0055] | [0.0047] | [0.0222] | [0.0110] | [0.0032] | [0.0019] | [0.0052] | [0.0024] | [0.0042] | [0.0045] |
| 27-Jun | Mean | **0.2684** | **0.3322** | 0.0086 | **0.1072** | **0.6744** | **0.6796** | **0.2684** | **0.3323** | 0.0085 | **0.1073** | **0.6745** | **0.6796** |
| | Std | [0.0029] | [0.0029] | [0.0046] | [0.0023] | [0.0040] | [0.0040] | [0.0032] | [0.0015] | [0.0061] | [0.0023] | [0.0041] | [0.0040] |
| 28-Jun | Mean | **0.2571** | **0.3190** | 0.0003 | **0.1126** | **0.6886** | **0.6757** | **0.2570** | **0.3191** | 0.0005 | **0.1126** | **0.6889** | **0.6759** |
| | Std | [0.0028] | [0.0028] | [0.0047] | [0.0023] | [0.0038] | [0.0040] | [0.0031] | [0.0014] | [0.0060] | [0.0023] | [0.0039] | [0.0041] |
| 29-Jun | Mean | **0.2597** | **0.3010** | **0.0179** | **0.0979** | **0.7011** | **0.6973** | **0.2598** | **0.3011** | **0.0179** | **0.0978** | **0.7013** | **0.6974** |
| | Std | [0.0028] | [0.0027] | [0.0043] | [0.0021] | [0.0036] | [0.0036] | [0.0030] | [0.0012] | [0.0057] | [0.0021] | [0.0036] | [0.0037] |
| Average | Mean | 0.2945 | 0.3298 | 0.0208 | 0.0986 | 0.6897 | 0.7079 | 0.2970 | 0.3350 | 0.0264 | 0.0982 | 0.6910 | 0.7086 |

## Table 7. The extended GH Model (GH) : MLE on 19 and 23 June 2016

$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t (q_t - E[q_t \mid \Omega_{t-1}]) + u_t, u_t \sim N(0, \sigma_u^2), q_t \in \{-1, 1\}$
where $\Pr[q_t = 1 \mid q_{t-1} = 1] = P, \Pr[q_t = -1 \mid q_{t-1} = -1] = Q$ and $\lambda_t = \lambda_0 + \lambda_1 \sqrt{V_t}$

This table provides the MLE estimates of the above model for every 10,000 trade using the gold futures contract traded on the CME Globex electronic trading platform on May 19, 2016 and June 23, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper.) $\sigma_u$ is the standard deviation of the (log) efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively, due to adverse selection. In this table, 10,000 fold of $c$, $\lambda_0$, $\lambda_1$, and $\sigma_u$ estimates are reported.

| Panel A. | | GH | | | | | |
|---|---|---|---|---|---|---|---|
| Intervals | 19.Jun | $c \times 10,000$ | $\sigma_u \times 10,000$ | $\lambda_0 \times 10,000$ | $\lambda_1 \times 10,000$ | $P$ | $Q$ |
| 1 | to 10,000 | **6.4691** | **1.1455** | 0.4026 | **0.2973** | **0.7779** | **0.9952** |
| | | [0.2488] | [0.0092] | [0.2663] | [0.0296] | [0.0285] | [0.0007] |
| 2 | to 20,000 | **0.2369** | **0.3296** | **0.0224** | **0.1079** | **0.6998** | **0.6812** |
| | | [0.0070] | [0.0067] | [0.0112] | [0.0045] | [0.0093] | [0.0103] |
| 3 | to 30,000 | **0.2618** | **0.3253** | -0.0045 | **0.1143** | **0.6932** | **0.6696** |
| | | [0.0067] | [0.0067] | [0.0107] | [0.0045] | [0.0084] | [0.0093] |
| 4 | to 40,000 | **3.8289** | **0.4982** | 0.1617 | 0.0132 | **0.9756** | **0.9978** |
| | | [0.3346] | [0.0047] | [0.3525] | [0.0140] | [0.0055] | [0.0005] |
| 5 | over 40,000 | **0.2338** | **0.3163** | **0.0279** | **0.1034** | **0.7129** | **0.7193** |
| | | [0.0056] | [0.0051] | [0.0088] | [0.0037] | [0.0070] | [0.0066] |
| | Average | **2.2061** | **0.5230** | 0.1220 | **0.1272** | **0.7719** | **0.8126** |
| | | [0.0726] | [0.0067] | [0.0910] | [0.0072] | [0.0083] | [0.0011] |

| Panel B. | | GH | | | | | |
|---|---|---|---|---|---|---|---|
| Intervals | 23.Jun | $c \times 10,000$ | $\sigma_u \times 10,000$ | $\lambda_0 \times 10,000$ | $\lambda_1 \times 10,000$ | $P$ | $Q$ |
| 1 | to 10,000 | **0.5888** | **0.8663** | 0.0013 | **0.2801** | **0.6134** | **0.8806** |
| | | [0.0379] | [0.0154] | [0.0413] | [0.0159] | [0.0293] | [0.0113] |
| 2 | to 20,000 | **0.4937** | **0.7023** | **-0.1802** | **0.3249** | **0.6875** | **0.7321** |
| | | [0.0178] | [0.0096] | [0.0266] | [0.0132] | [0.0224] | [0.0199] |
| 3 | to 30,000 | **0.5979** | **0.6508** | **-0.0602** | **0.2135** | **0.7514** | **0.6862** |
| | | [0.0160] | [0.0085] | [0.0237] | [0.0118] | [0.0121] | [0.0158] |
| 4 | to 40,000 | **0.7480** | **0.8385** | **-0.2228** | **0.3381** | **0.8047** | **0.6680** |
| | | [0.0229] | [0.0118] | [0.0331] | [0.0152] | [0.0117] | [0.0198] |
| 5 | to 50,000 | **0.6967** | **0.7770** | **-0.1401** | **0.3168** | **0.5469** | **0.7715** |
| | | [0.0161] | [0.0099] | [0.0247] | [0.0107] | [0.0126] | [0.0072] |
| 6 | to 60,000 | **1.9791** | **1.4917** | **-0.3999** | **0.5310** | **0.4938** | **0.9656** |
| | | [0.0767] | [0.0144] | [0.0944] | [0.0365] | [0.0215] | [0.0026] |
| 7 | to 70,000 | **1.8634** | **1.6498** | **-0.6026** | **0.7196** | **0.5066** | **0.9635** |
| | | [0.0993] | [0.0175] | [0.1156] | [0.0381] | [0.0268] | [0.0035] |
| 8 | to 80,000 | **0.9524** | **0.9531** | **-0.2908** | **0.4380** | **0.5595** | **0.8842** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | [0.0301] | [0.0120] | [0.0418] | [0.0199] | [0.0173] | [0.0061] |
| 9 | to 90,000 | **0.7291** | **0.8546** | **-0.3889** | **0.4027** | **0.3652** | **0.9482** |
| | | [0.0530] | [0.0115] | [0.0559] | [0.0250] | [0.0323] | [0.0077] |
| 10 | to 100,000 | **0.6063** | **0.8177** | **-0.2904** | **0.3642** | **0.9135** | **0.5138** |
| | | [0.0479] | [0.0141] | [0.0448] | [0.0216] | [0.0122] | [0.0281] |
| 11 | to 110,000 | **0.4745** | **0.6399** | **-0.0942** | **0.2284** | **0.6774** | **0.7772** |
| | | [0.0192] | [0.0100] | [0.0242] | [0.0110] | [0.0261] | [0.0203] |
| 12 | to 120,000 | **0.4426** | **0.6456** | **-0.1777** | **0.2620** | **0.8207** | **0.5577** |
| | | [0.0207] | [0.0108] | [0.0264] | [0.0133] | [0.0133] | [0.0259] |
| 13 | to 130,000 | **0.4085** | **0.5672** | **-0.2453** | **0.2824** | **0.9430** | **0.5707** |
| | | [0.0757] | [0.0111] | [0.0387] | [0.0199] | [0.0172] | [0.0281] |
| 14 | to 140,000 | **0.1929** | **0.4566** | **-0.1674** | **0.2202** | **0.6016** | **0.6904** |
| | | [0.0117] | [0.0071] | [0.0150] | [0.0070] | [0.0173] | [0.0168] |
| 15 | to 150,000 | **0.2419** | **0.3689** | **-0.0944** | **0.1850** | **0.6870** | **0.6956** |
| | | [0.0082] | [0.0069] | [0.0131] | [0.0066] | [0.0126] | [0.0114] |
| 16 | to 160,000 | **0.0879** | **0.4589** | **-0.2227** | **0.2584** | **0.4040** | **0.2585** |
| | | [0.0107] | [0.0055] | [0.0214] | [0.0085] | [0.0489] | [0.0313] |
| 17 | over 160,000 | **0.8761** | **0.5628** | -0.0550 | **0.1364** | **0.9928** | **0.5789** |
| | | [0.0719] | [0.0052] | [0.0754] | [0.0192] | [0.0013] | [0.0604] |
| | Average | **0.7047** | **0.7825** | **-0.2136** | **0.3236** | **0.6688** | **0.7143** |
| | | [0.0318] | [0.0106] | [0.0370] | [0.0156] | [0.0082] | [0.0085] |

## Table 8. Classified Trades: Correlations

This table provides daily correlation estimates of the classified trades using the Tick rule, the Roll model, and the extended GH model for the gold futures contract traded on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper). The daily correlations of classified trades between the Tick rule and the Roll model (the extended GH model) is labelled as Roll (GH) respectively.

| Date | Roll | GH |
|---|---|---|
| 1-May | 0.9999 | 0.9984 |
| 2-May | 1.0000 | 0.9991 |
| 3-May | 0.9980 | 0.9988 |
| 4-May | 0.9994 | 0.9995 |
| 5-May | 1.0000 | 0.9975 |
| 8-May | 0.9995 | 0.9998 |
| 9-May | 1.0000 | 0.9996 |
| 10-May | 1.0000 | 0.9998 |
| 11-May | 0.9999 | 0.9995 |
| 12-May | 1.0000 | 0.9994 |
| 15-May | 1.0000 | 0.9996 |
| 16-May | 1.0000 | 0.9999 |
| 17-May | 0.9996 | 0.9991 |
| 18-May | 1.0000 | 0.9999 |
| 19-May | 1.0000 | 0.9999 |
| 22-May | 0.9999 | 0.9998 |
| 23-May | 1.0000 | 1.0000 |
| 24-May | 1.0000 | 0.9990 |
| 25-May | 0.9999 | 1.0000 |
| 26-May | 0.9995 | 0.9997 |
| 29-May | 0.9998 | 0.9949 |
| 30-May | 1.0000 | 1.0000 |
| 31-May | 1.0000 | 1.0000 |
| 1-Jun | 1.0000 | 0.9999 |
| 2-Jun | 0.9998 | 0.9939 |
| 5-Jun | 1.0000 | 0.9998 |
| 6-Jun | 1.0000 | 1.0000 |
| 7-Jun | 1.0000 | 1.0000 |
| 8-Jun | 1.0000 | 0.9999 |
| 9-Jun | 1.0000 | 1.0000 |
| 12-Jun | 0.9995 | 0.9997 |
| 13-Jun | 1.0000 | 1.0000 |
| 14-Jun | 0.9995 | 0.9996 |
| 15-Jun | 1.0000 | 0.9997 |
| 16-Jun | 1.0000 | 0.9998 |
| 19-Jun | 0.9887 | 0.6063 |
| 20-Jun | 0.9995 | 0.9977 |
| 21-Jun | 0.9991 | 0.9960 |
| 22-Jun | 0.9989 | 0.9839 |
| 23-Jun | 0.9988 | 0.4455 |
| 26-Jun | 0.9991 | 0.8260 |
| 27-Jun | 0.9993 | 0.9976 |
| 28-Jun | 1.0000 | 0.9979 |
| 29-Jun | 0.9997 | 0.9996 |

## Table 9. The Bid-Ask Spread (ticks) and Relative Decomposition

$$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t\left(q_t - E\left[q_t \mid \Omega_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1,1\}$$

where $\Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\lambda_t = \lambda_0 + \lambda_1\sqrt{V_t}$

This table gives daily bid-ask spread in ticks and estimates of liquidity components of the above model with data from gold futures contracts traded on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper) In this model, $\sigma_u$ is the standard deviation of the log efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively. In particular, this table first reports $S_{GH,cs}$ which is the spread in ticks computed as $S_{GH,cs} = sp_{GH,cs} \times \overline{P}$ where $sp_{GH,cs}$ is the log spread estimates implied by the model with the daily average volume per trade($\overline{V_t}$) (i.e., computed as $sp_{GH,cs} = 2\times(c + \lambda_0 + \lambda_1\sqrt{\overline{V_t}})$ ) , and $\overline{P}$ is the average daily tick price. For example, the $S_{GH,cs}$ on 1, May is 1.1970 (i.e., $2\times(0.4615\times10^{-4})$ × (1,296.9/0.1) ) where 1,296.9 is the mean of the daily prices and 0.1 is the tick size on 1, May. In the last two columns, the estimates of informational and non-informational components of the bid-ask spread ( $S_{GH,cs}$ ) are computed using the results in Table 6 as follows.

Proportion of spread arising from order processing cost component : TC =$c$ / $(c + (\lambda_0 + \lambda_1\sqrt{\overline{V_t}}))$

Proportion of spread arising from the information asymmetry component: IC=$(\lambda_0 + \lambda_1\sqrt{\overline{V_t}})$ / $(c + (\lambda_0 + \lambda_1\sqrt{\overline{V_t}}))$

| Date | AVG. daily price | $\overline{V}$ | $sp_{GH,cs}$ | $S_{GH,cs}$ | TC | IC |
|---|---|---|---|---|---|---|
| 1-May | 1296.9 | 2.87 | 0.9229 | 1.1970 | 0.5254 | 0.4746 |
| 2-May | 1294.3 | 2.85 | 0.9312 | 1.2053 | 0.5749 | 0.4251 |
| 3-May | 1283.4 | 2.90 | 0.9457 | 1.2138 | 0.5751 | 0.4249 |
| 4-May | 1280.6 | 2.84 | 0.9288 | 1.1894 | 0.5593 | 0.4407 |
| 5-May | 1289.1 | 2.88 | 0.9471 | 1.2209 | 0.5719 | 0.4281 |
| 8-May | 1274.3 | 3.03 | 0.9081 | 1.1572 | 0.5751 | 0.4249 |
| 9-May | 1265.0 | 2.80 | 0.9288 | 1.1750 | 0.5262 | 0.4738 |
| 10-May | 1275.5 | 2.93 | 0.9217 | 1.1755 | 0.5749 | 0.4251 |
| 11-May | 1273.2 | 2.93 | 0.9274 | 1.1808 | 0.5872 | 0.4128 |
| 12-May | 1270.9 | 2.82 | 0.9419 | 1.1971 | 0.5486 | 0.4514 |
| 15-May | 1281.0 | 3.06 | 0.9330 | 1.1952 | 0.6201 | 0.3799 |
| 16-May | 1277.5 | 2.96 | 0.9294 | 1.1873 | 0.5792 | 0.4208 |
| 17-May | 1271.3 | 2.93 | 0.9656 | 1.2275 | 0.5768 | 0.4232 |
| 18-May | 1252.7 | 3.32 | 0.9545 | 1.1956 | 0.5948 | 0.4052 |
| 19-May | 1255.3 | 2.95 | 0.9327 | 1.1708 | 0.5960 | 0.4040 |
| 22-May | 1249.8 | 2.82 | 0.9038 | 1.1296 | 0.6159 | 0.3841 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 23-May | 1237.6 | 3.34 | 0.9385 | 1.1615 | 0.6601 | 0.3399 |
| 24-May | 1223.2 | 3.16 | 0.9197 | 1.1250 | 0.6512 | 0.3488 |
| 25-May | 1226.5 | 2.71 | 0.9481 | 1.1627 | 0.5871 | 0.4129 |
| 26-May | 1216.6 | 3.18 | 0.9614 | 1.1320 | 0.5909 | 0.4091 |
| 29-May | 1207.7 | 2.82 | 1.0215 | 1.2336 | 0.5448 | 0.4552 |
| 30-May | 1215.3 | 2.96 | 0.9551 | 1.1608 | 0.6037 | 0.3963 |
| 31-May | 1216.4 | 2.99 | 0.9518 | 1.1577 | 0.6420 | 0.3580 |
| 1-Jun | 1215.7 | 2.70 | 0.9601 | 1.1672 | 0.5919 | 0.4081 |
| 2-Jun | 1234.1 | 3.30 | 0.9830 | 1.2132 | 0.6974 | 0.3026 |
| 5-Jun | 1246.2 | 3.05 | 0.9357 | 1.1660 | 0.6101 | 0.3899 |
| 6-Jun | 1243.9 | 3.02 | 0.9068 | 1.1279 | 0.6582 | 0.3418 |
| 7-Jun | 1258.8 | 3.39 | 0.8941 | 1.1254 | 0.6527 | 0.3473 |
| 8-Jun | 1267.0 | 3.09 | 0.8917 | 1.1298 | 0.6234 | 0.3766 |
| 9-Jun | 1273.8 | 3.13 | 0.8936 | 1.1383 | 0.6131 | 0.3869 |
| 12-Jun | 1284.0 | 3.19 | 0.8893 | 1.1419 | 0.6081 | 0.3919 |
| 13-Jun | 1286.5 | 3.14 | 0.8607 | 1.1073 | 0.6525 | 0.3475 |
| 14-Jun | 1289.6 | 3.01 | 0.9358 | 1.2068 | 0.6057 | 0.3943 |
| 15-Jun | 1305.1 | 3.17 | 0.8965 | 1.1699 | 0.6256 | 0.3744 |
| 16-Jun | 1290.6 | 3.07 | 0.8832 | 1.1398 | 0.6034 | 0.3966 |
| 19-Jun | 1287.6 | 3.18 | 4.8659 | 6.2654 | 0.9068 | 0.0932 |
| 20-Jun | 1278.5 | 3.03 | 0.9329 | 1.1927 | 0.5469 | 0.4531 |
| 21-Jun | 1269.4 | 2.84 | 0.9245 | 1.1735 | 0.5312 | 0.4688 |
| 22-Jun | 1266.0 | 2.69 | 0.9600 | 1.2154 | 0.5140 | 0.4860 |
| 23-Jun | 1315.0 | 2.88 | 2.0815 | 3.2990 | 0.6771 | 0.3229 |
| 26-Jun | 1330.2 | 2.83 | 0.8891 | 1.1827 | 0.6206 | 0.3794 |
| 27-Jun | 1317.7 | 2.96 | 0.9061 | 1.1940 | 0.5925 | 0.4075 |
| 28-Jun | 1323.8 | 3.01 | 0.9052 | 1.1982 | 0.5681 | 0.4319 |
| 29-Jun | 1320.8 | 3.03 | 0.8958 | 1.1359 | 0.5799 | 0.4201 |

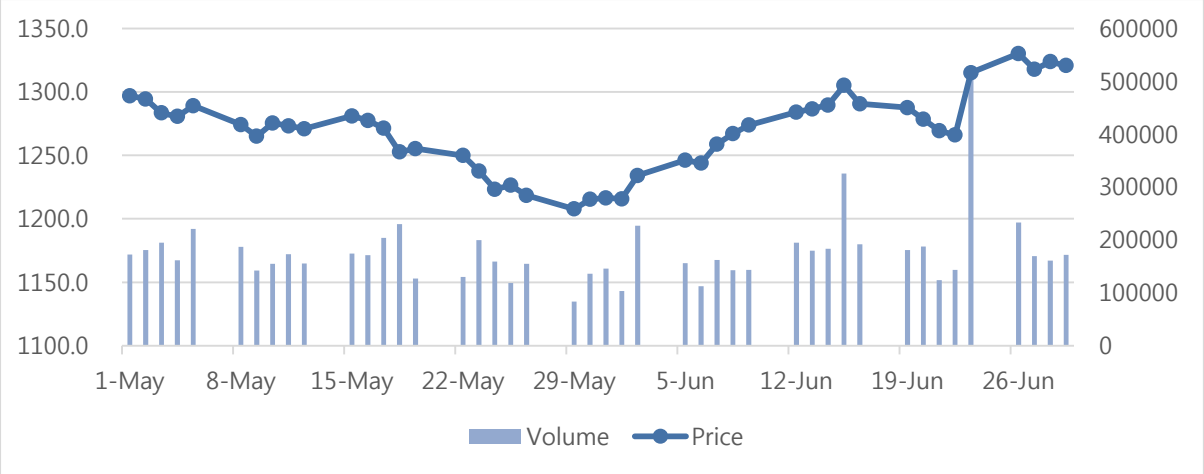


**Figure 1: Daily series of average prices (Price) and trading volumes (Volume)**

This figure presents the daily series of average prices (Price) and trading volumes (Volume) of the gold futures contract from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper)

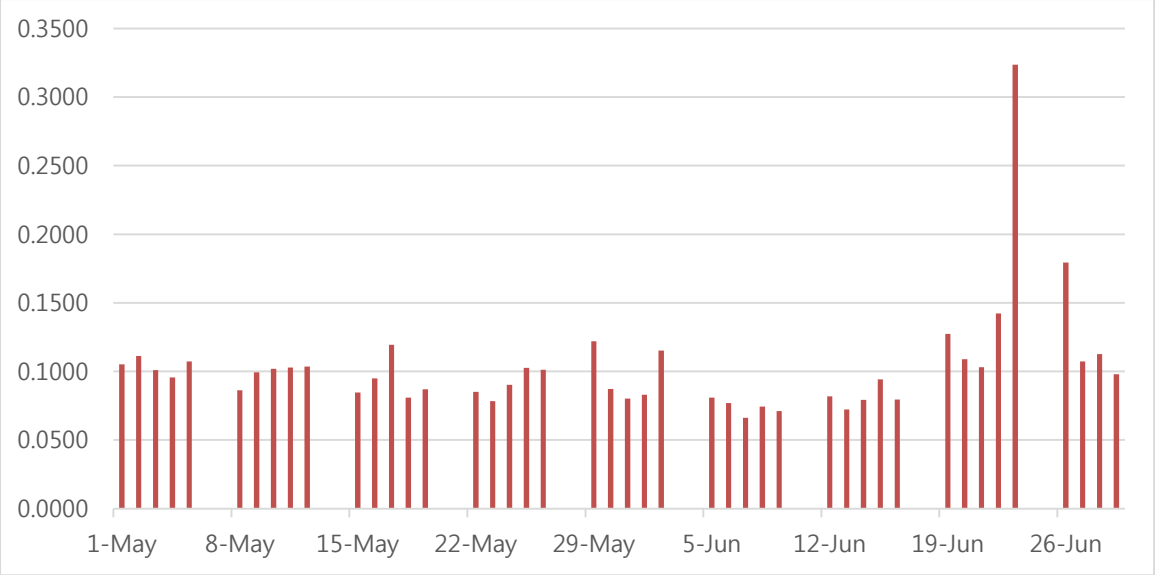


**Figure 2: Daily estimates of the permanent price impact**

$$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t \left(q_t - E\left[q_t \mid \Omega_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1, 1\}$$

where $\Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\lambda_t = \lambda_0 + \lambda_1 \sqrt{V_t}$

This figure provides daily $\lambda_1$ estimates from the above model using data on gold futures contracts trading on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper.) In this model, $\sigma_u$ is the standard deviation of the log efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively.

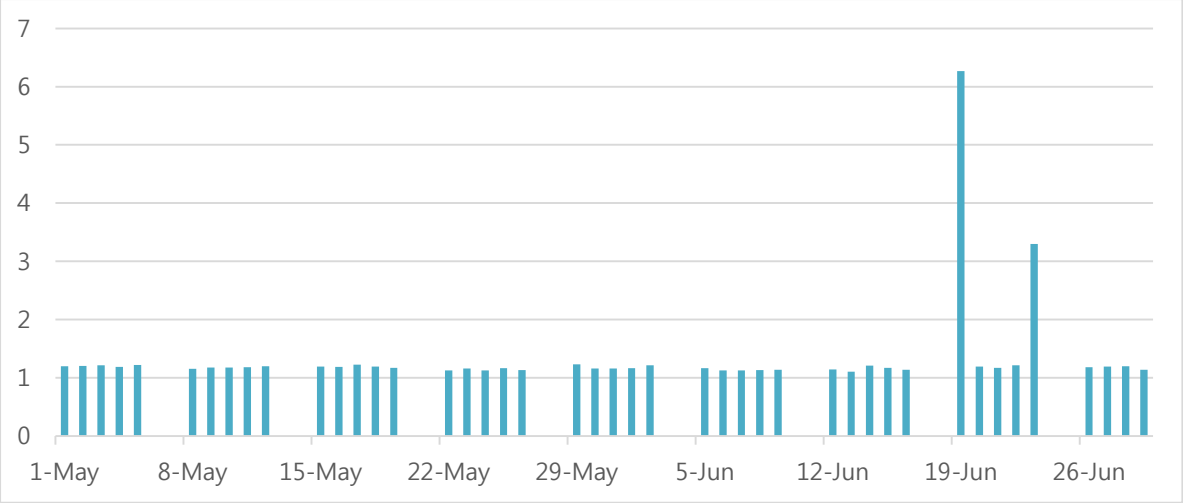


**Figure 3: The Bid-Ask Spread in ticks**

$$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t\left(q_t - E\left[q_t \mid \Omega_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1, 1\}$$

where $\Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\lambda_t = \lambda_0 + \lambda_1\sqrt{V_t}$

This figure presents daily bid-ask spread in ticks of the above model using data from gold futures contracts trading on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016. (See the caption of Table 2 for further details on the dating convention used in the paper.) In this model, $\sigma_u$ is the standard deviation of the log efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively. In particular, this table first reports $S_{GH,cs}$ which is the spread in ticks computed as $S_{GH,cs} = sp_{GH,cs} \times \overline{P}$ where $s_{GH,cs}$ is the log spread estimates implied by the model with the daily average volume ($\overline{V_t}$) (i.e., computed as $sp_{GH,cs} = 2 \times (c + \lambda_0 + \lambda_1\sqrt{\overline{V_t}})$ ), and $\overline{P}$ is the average daily tick price. For example, the $S_{GH,cs}$ on 1, May is 1.1970 (i.e., $2 \times (0.4615 \times 10^{-4}) \times (1{,}296.9/0.1)$) where 1,296.9 is the mean of the daily prices and 0.1 is the tick size on 1, May.

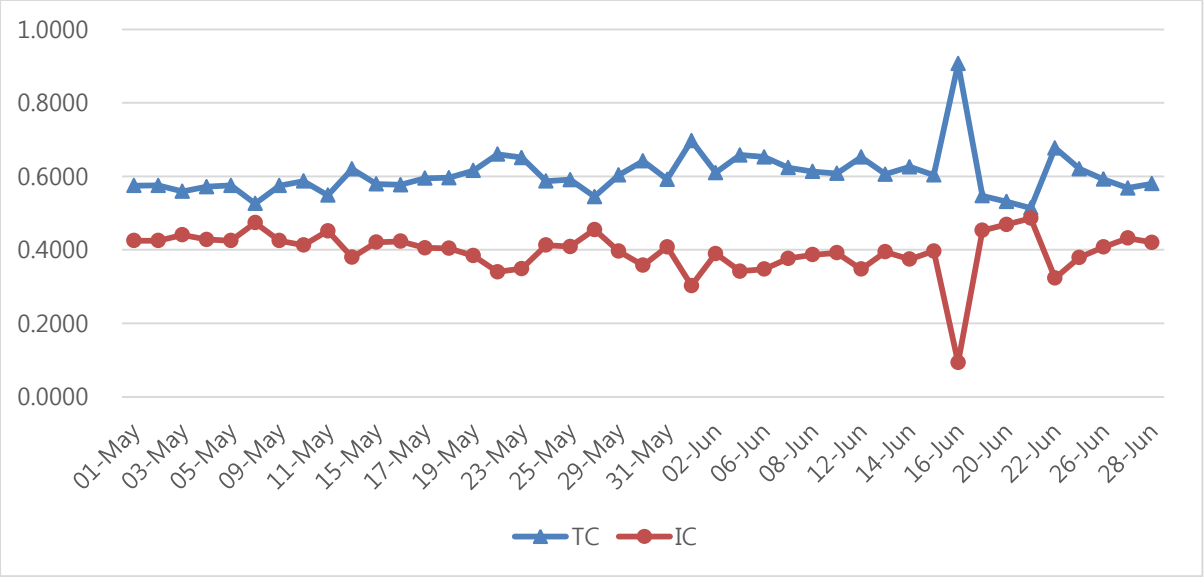


**Figure 4: Daily estimates of informational (IC) and non-informational (TC) components**

$$p_t = m_t + cq_t, m_t = m_{t-1} + \lambda_t \left(q_t - E\left[q_t \mid \Omega_{t-1}\right]\right) + u_t, u_t \sim N\left(0, \sigma_u^2\right), q_t \in \{-1, 1\}$$

where $\Pr\left[q_t = 1 \mid q_{t-1} = 1\right] = P, \Pr\left[q_t = -1 \mid q_{t-1} = -1\right] = Q$ and $\lambda_t = \lambda_0 + \lambda_1 \sqrt{V_t}$

This figure presents daily estimates of liquidity components of the above model using gold futures data from contracts trading on the CME Globex electronic trading platform from May 1, 2016 to June 30, 2016 (See the caption of Table 2 for further details on the dating convention used in the paper.) In this model, $\sigma_u$ is the standard deviation of the log efficient price changes: $c$ is the order processing component of the (log) half spread. $P$ and $Q$ are transition probabilities. $\lambda_0$ and $\lambda_1$ are the fixed and variable permanent price impact costs, respectively. In particular, the estimates of informational and non-informational components of the bid-ask spread ($S_{GH,cs}$) are computed using the results in Table 6 as follows.

Proportion of spread arising from order processing cost component: TC $= c / (c + (\lambda_0 + \lambda_1 \sqrt{\overline{V}_t}))$

Proportion of spread arising from the information asymmetry component: IC$= (\lambda_0 + \lambda_1 \sqrt{\overline{V}_t}) / (c + (\lambda_0 + \lambda_1 \sqrt{\overline{V}_t}))$

# APPENDICES

## Table A1. Convergence diagnostics for the Bayesian estimation

This table provides two convergence diagnostics for the Bayesian estimation of the three models: Roll, the extended Roll, and the extended GH. In particular, we compute and report the effective sample size based on the inefficiency factors and the p-value of Geweke(1992)'s convergence diagnostics test for the model parameters.

1) **Geweke(1992)'s convergence diagnostic:** if Geweke's p-value is less than 0.05, we interpret the burn-in period is too small and we can't guarantee the convergence of the chain. In general, as Geweke's p-value is close to 1, we have more efficient samples.
2) **Inefficiency factor and effective sample size: t**he effective sample size should be larger than 1000 for all parameters to guarantee sufficient number of the MCMC draws.

Refer to Appendix B for more details on the inefficiency factors and the Geweke(1992)'s diagnostic tests.

| | | Roll (Bayesian) | | MS(Bayeisan) | | | | GH(Bayeisan) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Date | Parameters | $c$ | $\sigma_u$ | $c$ | $\sigma_u$ | $P$ | $Q$ | $c$ | $\sigma_u$ | $\lambda_0$ | $\lambda_1$ | $P$ | $Q$ |
| 1-May | Geweke P | 0.9274 | 0.7340 | 0.7009 | 0.4356 | 0.5952 | 0.6594 | 0.5929 | 0.7866 | 0.5419 | 0.8123 | 0.6971 | 0.1978 |
| | Effective size | 8413 | 8998 | 11709 | 9239 | 1761 | 1809 | 9405 | 6301 | 5210 | 9793 | 3988 | 3772 |
| 2-May | Geweke P | 0.7120 | 0.8444 | 0.7632 | 0.7944 | 0.9771 | 0.6380 | 0.9897 | 0.9338 | 0.9901 | 0.9178 | 0.3291 | 0.9217 |
| | Effective size | 6693 | 6940 | 13726 | 13494 | 2203 | 2292 | 10202 | 8163 | 5358 | 11294 | 4305 | 4866 |
| 3-May | Geweke P | 0.9605 | 0.9132 | 0.8327 | 0.7237 | 0.9632 | 0.9860 | 0.6982 | 0.9823 | 0.8535 | 0.9734 | 0.7017 | 0.8055 |
| | Effective size | 8322 | 8432 | 13122 | 11211 | 1586 | 1615 | 10009 | 7062 | 5094 | 7955 | 3083 | 3163 |
| 4-May | Geweke P | 0.7671 | 0.7688 | 0.9416 | 0.7742 | 0.8531 | 0.8939 | 0.5546 | 0.4763 | 0.7750 | 0.9529 | 0.7296 | 0.1347 |
| | Effective size | 7091 | 7830 | 14005 | 13976 | 2098 | 2068 | 8863 | 7376 | 4871 | 10748 | 3995 | 3889 |
| 5-May | Geweke P | 0.9031 | 0.8974 | 0.5375 | 0.6463 | 0.1675 | 0.3989 | 0.5032 | 0.0945 | 0.3977 | 0.8397 | 0.7748 | 0.7234 |
| | Effective size | 8556 | 9566 | 12791 | 8561 | 1410 | 1440 | 10337 | 6100 | 5620 | 10743 | 3702 | 3730 |
| 8-May | Geweke P | 0.7246 | 0.9435 | 0.9898 | 0.8651 | 0.7313 | 0.8059 | 0.3239 | 0.6552 | 0.6819 | 0.8941 | 0.8464 | 0.8006 |
| | Effective size | 7552 | 8164 | 15635 | 16948 | 2039 | 2308 | 6978 | 9673 | 4153 | 8696 | 3292 | 4265 |
| 9-May | Geweke P | 0.9923 | 0.9756 | 0.6797 | 0.5197 | 0.6723 | 0.7198 | 0.9466 | 0.6883 | 0.7881 | 0.9219 | 0.8683 | 0.8723 |
| | Effective size | 7312 | 8065 | 13524 | 16028 | 2228 | 2239 | 9148 | 9055 | 4902 | 10650 | 4548 | 4300 |
| 10-May | Geweke P | 0.8597 | 0.8468 | 0.9085 | 0.9910 | 0.8637 | 0.8849 | 0.9912 | 0.9483 | 0.8477 | 0.9829 | 0.5173 | 0.5464 |
| | Effective size | 4355 | 4450 | 16384 | 22530 | 2478 | 2411 | 9797 | 10977 | 5277 | 11881 | 4414 | 4630 |
| 11-May | Geweke P | 0.7258 | 0.7233 | 0.9749 | 0.9444 | 0.7256 | 0.7679 | 0.9873 | 0.8992 | 0.6461 | 0.9855 | 0.8663 | 0.7997 |
| | Effective size | 6011 | 6048 | 15662 | 16480 | 2442 | 2521 | 9102 | 9629 | 5154 | 9822 | 4496 | 4650 |
| 12-May | Geweke P | 0.4671 | 0.1306 | 0.8781 | 0.8194 | 0.5045 | 0.4020 | 0.8810 | 0.8251 | 0.4436 | 0.6628 | 0.7698 | 0.8778 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Effective size | 5998 | 6269 | 15485 | 14156 | 2336 | 2521 | 10673 | 8176 | 5254 | 10002 | 4387 | 5525 |
| 15-May | Geweke P | 0.9804 | 0.9910 | 0.7367 | 0.9856 | 0.2007 | 0.8215 | 0.5584 | 0.6684 | 0.7590 | 0.8321 | 0.9885 | 0.9578 |
| | Effective size | 1110 | 1104 | 19063 | 22743 | 2440 | 2581 | 9392 | 9749 | 5663 | 14534 | 4491 | 4378 |
| 16-May | Geweke P | 0.9861 | 0.9982 | 0.9812 | 0.8226 | 0.8437 | 0.9154 | 0.9140 | 0.6349 | 0.9917 | 0.8659 | 0.9325 | 0.9850 |
| | Effective size | 4731 | 4964 | 16268 | 21081 | 2479 | 2479 | 10010 | 10550 | 5697 | 11989 | 4254 | 4262 |
| 17-May | Geweke P | 0.9408 | 0.9837 | 0.9529 | 0.9817 | 0.8258 | 0.7747 | 0.9903 | 0.7454 | 0.8285 | 0.6952 | 0.9203 | 0.9801 |
| | Effective size | 6925 | 7316 | 13956 | 11694 | 1777 | 1823 | 9781 | 7233 | 5551 | 9936 | 3902 | 4169 |
| 18-May | Geweke P | 0.9808 | 0.9915 | 0.7779 | 0.8919 | 0.5869 | 0.5827 | 0.9980 | 0.9564 | 0.9744 | 0.9993 | 0.9060 | 0.7648 |
| | Effective size | 2755 | 2765 | 18418 | 23314 | 2695 | 2865 | 22352 | 25950 | 47455 | 170478 | 10429 | 12956 |
| 19-May | Geweke P | 0.8704 | 0.8706 | 0.9976 | 0.7876 | 0.9471 | 0.9925 | 0.8460 | 0.9024 | 0.7672 | 0.7379 | 0.8028 | 0.6164 |
| | Effective size | 1915 | 1920 | 17516 | 25642 | 2995 | 3043 | 9078 | 11419 | 5420 | 14514 | 4634 | 5869 |
| 22-May | Geweke P | 0.8781 | 0.8970 | 0.9675 | 0.9082 | 0.9316 | 0.9432 | 0.7572 | 0.9166 | 0.9321 | 0.2853 | 0.6214 | 0.5764 |
| | Effective size | 2498 | 2554 | 16569 | 33538 | 2555 | 2635 | 6881 | 12558 | 4287 | 12419 | 3653 | 4609 |
| 23-May | Geweke P | 0.7080 | 0.8577 | 0.8650 | 0.8628 | 0.4451 | 0.0017 | 0.8091 | 0.9661 | 0.9791 | 0.9982 | 0.9101 | 0.9036 |
| | Effective size | 10768 | 10680 | 18213 | 39618 | 1970 | 2016 | 7423 | 15025 | 4428 | 11320 | 3793 | 3681 |
| 24-May | Geweke P | 0.4406 | 0.6900 | 0.7203 | 0.8073 | 0.9700 | 0.8814 | 0.6428 | 0.7684 | 0.7129 | 0.7195 | 0.8326 | 0.4196 |
| | Effective size | 9734 | 8897 | 17721 | 40108 | 2246 | 2086 | 8064 | 22279 | 4768 | 9563 | 4227 | 4033 |
| 25-May | Geweke P | 0.8894 | 0.8665 | 0.9421 | 0.8548 | 0.7626 | 0.7971 | 0.9654 | 0.9888 | 0.8617 | 0.6404 | 0.9288 | 0.9292 |
| | Effective size | 6131 | 6708 | 18609 | 20088 | 2407 | 2504 | 8048 | 9792 | 4614 | 8849 | 3835 | 4483 |
| 26-May | Geweke P | 0.4039 | 0.8067 | 0.9523 | 0.9749 | 0.9754 | 0.9128 | 0.2232 | 0.6810 | 0.4580 | 0.6544 | 0.3993 | 0.4403 |
| | Effective size | 5997 | 6121 | 17895 | 21393 | 2468 | 2548 | 10901 | 13522 | 5272 | 8721 | 4332 | 4447 |
| 29-May | Geweke P | 0.8549 | 0.8375 | 0.9162 | 0.9685 | 0.7571 | 0.1063 | 0.7973 | 0.5983 | 0.6381 | 0.8937 | 0.5263 | 0.0568 |
| | Effective size | 7427 | 7657 | 11083 | 9611 | 2219 | 2498 | 10980 | 6422 | 6110 | 11504 | 3862 | 5104 |
| 30-May | Geweke P | 0.8045 | 0.8141 | 0.7647 | 0.9484 | 0.7140 | 0.6880 | 0.6616 | 0.9484 | 0.8442 | 0.9858 | 0.4684 | 0.9291 |
| | Effective size | 3001 | 2921 | 19121 | 28204 | 3115 | 3053 | 8867 | 11798 | 5117 | 13621 | 4875 | 4487 |
| 31-May | Geweke P | 0.8063 | 0.7697 | 0.5694 | 0.9738 | 0.6201 | 0.7709 | 0.6797 | 0.7483 | 0.1615 | 0.4906 | 0.8193 | 0.8537 |
| | Effective size | 11536 | 10198 | 22675 | 40217 | 3217 | 3210 | 8190 | 12667 | 5347 | 14808 | 4416 | 4129 |
| 1-Jun | Geweke P | 0.9828 | 0.9945 | 0.9642 | 0.6798 | 0.9261 | 0.9397 | 0.5620 | 0.6319 | 0.2861 | 0.7874 | 0.4802 | 0.9943 |
| | Effective size | 1921 | 1920 | 16855 | 23509 | 3054 | 2898 | 8539 | 10738 | 5241 | 11107 | 4903 | 4214 |
| 2-Jun | Geweke P | 0.9627 | 0.8922 | 0.8057 | 0.6785 | 0.8385 | 0.7904 | 0.4829 | 0.6432 | 0.5624 | 0.1216 | 0.7074 | 0.9768 |
| | Effective size | 9166 | 10011 | 2100 | 3996 | 1593 | 2303 | 8562 | 8948 | 5794 | 16780 | 3924 | 3970 |
| 5-Jun | Geweke P | 0.7646 | 0.7154 | 0.9174 | 0.8756 | 0.9172 | 0.8797 | 0.8476 | 0.7030 | 0.2764 | 0.6942 | 0.5404 | 0.8952 |
| | Effective size | 4603 | 4810 | 17471 | 25441 | 2046 | 2098 | 7583 | 10716 | 4823 | 12453 | 3758 | 4054 |
| 6-Jun | Geweke P | 0.9076 | 0.8397 | 0.8185 | 0.9313 | 0.9465 | 0.8586 | 0.9260 | 0.5150 | 0.7924 | 0.7595 | 0.9570 | 0.9982 |
| | Effective size | 14111 | 12035 | 20441 | 46255 | 3599 | 3606 | 7152 | 14816 | 4388 | 11504 | 4585 | 3779 |
| 7-Jun | Geweke P | 0.9111 | 0.9243 | 0.3407 | 0.9785 | 0.6226 | 0.7852 | 0.8492 | 0.7661 | 0.7781 | 0.6150 | 0.3032 | 0.6276 |
| | Effective size | 12290 | 11849 | 20534 | 44770 | 2900 | 2987 | 6984 | 16289 | 4173 | 13000 | 4067 | 4286 |
| 8-Jun | Geweke P | 0.9028 | 0.7456 | 0.9511 | 0.9431 | 0.6865 | 0.6674 | 0.6449 | 0.6903 | 0.3703 | 0.7814 | 0.9207 | 0.9178 |
| | Effective size | 9966 | 9624 | 20373 | 41532 | 3583 | 3449 | 8571 | 15989 | 5009 | 12590 | 4736 | 4527 |
| 9-Jun | Geweke P | 0.6104 | 0.4483 | 0.9445 | 0.9229 | 0.9531 | 0.9779 | 0.6278 | 0.8647 | 0.1998 | 0.8142 | 0.6720 | 0.1087 |
| | Effective size | 6233 | 6095 | 17409 | 34046 | 3001 | 2882 | 8419 | 13675 | 4854 | 12705 | 4078 | 3924 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12-Jun | Geweke P | 0.8979 | 0.8997 | 0.8840 | 0.7422 | 0.5844 | 0.8830 | 0.8565 | 0.3687 | 0.5347 | 0.6629 | 0.8213 | 0.5067 |
| | Effective size | 4340 | 4408 | 16462 | 24304 | 2508 | 2432 | 8792 | 12149 | 4823 | 11415 | 3772 | 3205 |
| 13-Jun | Geweke P | 0.2990 | 0.7920 | 0.9485 | 0.9339 | 0.9445 | 0.8870 | 0.9135 | 0.9903 | 0.9949 | 0.9700 | 0.7857 | 0.8169 |
| | Effective size | 10497 | 9993 | 18318 | 44846 | 2824 | 2754 | 7583 | 19914 | 4421 | 11163 | 3725 | 3829 |
| 14-Jun | Geweke P | 0.9301 | 0.9365 | 0.7596 | 0.8193 | 0.7304 | 0.7161 | 0.8014 | 0.9051 | 0.8559 | 0.9507 | 0.9008 | 0.7635 |
| | Effective size | 7551 | 7924 | 14613 | 11401 | 1299 | 1278 | 7698 | 7484 | 5386 | 11745 | 2985 | 2777 |
| 15-Jun | Geweke P | 0.9551 | 0.9655 | 0.9315 | 0.9434 | 0.8175 | 0.7730 | 0.9280 | 0.8710 | 0.9975 | 0.9427 | 0.6644 | 0.6701 |
| | Effective size | 3767 | 3785 | 19552 | 24541 | 2649 | 2604 | 6351 | 12261 | 3669 | 6334 | 3840 | 3211 |
| 16-Jun | Geweke P | 0.9779 | 0.9994 | 0.9391 | 0.9648 | 0.5655 | 0.5579 | 0.7463 | 0.7460 | 0.7892 | 0.7972 | 0.5594 | 0.9741 |
| | Effective size | 3350 | 3386 | 17458 | 28623 | 2255 | 2355 | 8194 | 12251 | 4583 | 11865 | 4021 | 4214 |
| 19-Jun | Geweke P | 0.9612 | 0.9813 | 0.3395 | 0.3643 | 0.1126 | 0.3096 | 0.3752 | 0.4261 | 0.4189 | 0.3922 | 0.3181 | 0.3925 |
| | Effective size | 10292 | 12942 | 1795 | 5090 | 1443 | 2416 | 1761 | 1855 | 1800 | 1754 | 1336 | 1130 |
| 20-Jun | Geweke P | 0.9796 | 0.9958 | 0.9473 | 0.5442 | 0.9371 | 0.8569 | 0.8620 | 0.9588 | 0.9405 | 0.9189 | 0.9376 | 0.9126 |
| | Effective size | 8563 | 9767 | 10851 | 9542 | 1813 | 1761 | 10706 | 7066 | 5464 | 8889 | 3483 | 3194 |
| 21-Jun | Geweke P | 0.6503 | 0.8400 | 0.9390 | 0.7055 | 0.7626 | 0.8244 | 0.8835 | 0.7967 | 0.8501 | 0.4754 | 0.7167 | 0.9325 |
| | Effective size | 8429 | 9114 | 12092 | 11998 | 2339 | 2167 | 10139 | 8849 | 5454 | 10399 | 4308 | 3487 |
| 22-Jun | Geweke P | 0.7910 | 0.9317 | 0.8806 | 0.9841 | 0.8713 | 0.8793 | 0.9420 | 0.8065 | 0.8815 | 0.9087 | 0.7090 | 0.7914 |
| | Effective size | 8430 | 9672 | 8324 | 5570 | 1102 | 1119 | 2765 | 2468 | 1689 | 2426 | 2231 | 2358 |
| 23-Jun | Geweke P | 0.9840 | 0.9498 | 0.9698 | 0.2067 | 0.0281 | 0.3924 | 0.5930 | 0.8584 | 0.8570 | 0.8807 | 0.3825 | 0.7881 |
| | Effective size | 12473 | 16675 | 3445 | 3441 | 1404 | 3537 | 2175 | 3563 | 2160 | 2260 | 1533 | 2951 |
| 26-Jun | Geweke P | 0.9672 | 0.9994 | 0.9156 | 0.7016 | 0.9716 | 0.3641 | 0.8711 | 0.7982 | 0.9228 | 0.7864 | 0.4213 | 0.3246 |
| | Effective size | 10103 | 12038 | 2206 | 3219 | 2184 | 1168 | 1272 | 2429 | 1068 | 1255 | 2002 | 1894 |
| 27-Jun | Geweke P | 0.7884 | 0.7695 | 0.9528 | 0.9358 | 0.9717 | 0.9076 | 0.7976 | 0.5119 | 0.5352 | 0.7927 | 0.5152 | 0.3013 |
| | Effective size | 8078 | 8351 | 12859 | 8734 | 1761 | 1914 | 9493 | 6378 | 5683 | 11834 | 3190 | 3189 |
| 28-Jun | Geweke P | 0.7134 | 0.9336 | 0.9197 | 0.7476 | 0.3981 | 0.3297 | 0.7854 | 0.9133 | 0.9032 | 0.9169 | 0.8185 | 0.8430 |
| | Effective size | 7551 | 7955 | 13293 | 11938 | 1875 | 1817 | 8604 | 6969 | 4565 | 8044 | 4115 | 3649 |
| 29-Jun | Geweke P | 0.7966 | 0.6836 | 0.9112 | 0.8719 | 0.8441 | 0.9496 | 0.8172 | 0.8332 | 0.8920 | 0.8856 | 0.7177 | 0.7762 |
| | Effective size | 7797 | 8246 | 14997 | 15705 | 2072 | 2067 | 9272 | 9592 | 5147 | 9459 | 4436 | 4205 |

## Table A.2 The difference of estimates from Bayes and MLE

This table presents average and maximum absolute differences between parameter estimates based on MLE and Bayesian estimates from the Roll, the extended Roll (MS), and the extended GH (GH) model.

| Bayes - MLE | parameters | Roll | MS | GH |
|---|---|---|---|---|
| Average difference | $c \times 10{,}000$ | 0.0000 | 0.0001 | -0.0010 |
| | $\sigma_u \times 10{,}000$ | 0.0000 | 0.0000 | -0.0027 |
| | $\lambda_0 \times 10{,}000$ | - | - | 0.0052 |
| | $\lambda_1 \times 10{,}000$ | - | - | -0.0021 |
| | $P$ | - | 0.0003 | -0.0021 |
| | $Q$ | - | 0.0000 | 0.0009 |

| Bayes - MLE | Parameters | Roll | MS | GH |
|---|---|---|---|---|
| Max abs(difference) | $c \times 10{,}000$ | 0.0013 | 0.0019 | 0.0272 |
| | $\sigma_u \times 10{,}000$ | 0.0013 | 0.0004 | 0.0702 |
| | $\lambda_0 \times 10{,}000$ | - | - | 0.1223 |
| | $\lambda_1 \times 10{,}000$ | - | - | 0.0513 |
| | $P$ | - | 0.0074 | 0.0845 |
| | $Q$ | - | 0.0013 | 0.0381 |

## Table A3. Model comparison (Likelihood ratio test)

This table presents model comparison results from the Roll, and the extended Roll (MS), and the extended GH (GH) models based on the likelihood ratio tests as follows.

$$-2\ln\left(\frac{\hat{L}_R}{\hat{L}_U}\right) \sim \chi_k^2$$

where $\hat{L}_R$ and $\hat{L}_U$ are the likelihood value under restricted model and unrestricted model, respectively. For example, when we compare the Roll model with the extended GH model, $\hat{L}_R$ and $\hat{L}_U$ are the likelihood value under the Roll model and under the extended GH model, respectively. Since we need to put restrictions on four parameters ( $P$ , $Q$ , $\lambda_0$ and $\lambda_1$ ) on the extended GH model to obtain the Roll model, the test statistics follows a chi-square distribution with 4 degrees of freedom.

Crit(1%), 6.6349 (Roll-MS), 11.3449 (Roll-GH), 9.2103 (MS-GH)

| Date | Model | Likelihood | Likelihood ratio test | | | |
|---|---|---|---|---|---|---|
| 1-May | Roll | 506516.3411 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 507429.8096 | Chi-stat | 1826.9369 | 4544.4196 | 2717.4827 |
| | GH | 508788.5509 | P-value | 0 | 0 | 0 |
| 2-May | Roll | 534907.565 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 536174.3543 | Chi-stat | 2533.5785 | 5993.3090 | 3459.7305 |
| | GH | 537904.2195 | P-value | 0 | 0 | 0 |
| 3-May | Roll | 563875.5491 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 565407.3009 | Chi-stat | 3063.5037 | 6233.1780 | 3169.6743 |
| | GH | 566992.1381 | P-value | 0 | 0 | 0 |
| 4-May | Roll | 480310.5331 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 481533.6496 | Chi-stat | 2446.2331 | 4987.4244 | 2541.1914 |
| | GH | 482804.2453 | P-value | 0 | 0 | 0 |
| 5-May | Roll | 643872.5622 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 645299.3172 | Chi-stat | 2853.5099 | 6354.5856 | 3501.0757 |
| | GH | 647049.855 | P-value | 0 | 0 | 0 |
| 8-May | Roll | 521049.5271 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 523056.4188 | Chi-stat | 4013.7834 | 6790.0208 | 2776.2374 |
| | GH | 524444.5375 | P-value | 0 | 0 | 0 |
| 9-May | Roll | 429310.7187 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 430505.897 | Chi-stat | 2390.3566 | 4884.1660 | 2493.8094 |
| | GH | 431752.8017 | P-value | 0 | 0 | 0 |
| 10-May | Roll | 445413.9822 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 447043.2221 | Chi-stat | 3258.4798 | 6198.8750 | 2940.3953 |
| | GH | 448513.4197 | P-value | 0 | 0 | 0 |
| 11-May | Roll | 498291.7839 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 499815.2659 | Chi-stat | 3046.9641 | 5984.7968 | 2937.8328 |
| | GH | 501284.1823 | P-value | 0 | 0 | 0 |
| 12-May | Roll | 463180.7376 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 464474.9132 | Chi-stat | 2588.3511 | 5352.4716 | 2764.1205 |
| | GH | 465856.9734 | P-value | 0 | 0 | 0 |
| 15-May | Roll | 479958.3703 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 481619.7409 | Chi-stat | 3322.7412 | 5761.2219 | 2438.4807 |
| | GH | 482838.9813 | P-value | 0 | 0 | 0 |
| 16-May | Roll | 488362.0101 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |

| | | | | | | |
|---|---|---|---|---|---|---|
| | MS | 489909.8219 | Chi-stat | 3095.6235 | 5825.5533 | 2729.9298 |
| | GH | 491274.7868 | P-value | 0 | 0 | 0 |
| 17-May | Roll | 582443.3422 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 583638.2658 | Chi-stat | 2389.8472 | 6063.9801 | 3674.1329 |
| | GH | 585475.3323 | P-value | 0 | 0 | 0 |
| 18-May | Roll | 581842.2541 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 584348.0668 | Chi-stat | 5011.6255 | 8264.8027 | 3253.1771 |
| | GH | 585974.6554 | P-value | 0 | 0 | 0 |
| 19-May | Roll | 362660.0529 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 364184.207 | Chi-stat | 3048.3083 | 4902.4966 | 1854.1883 |
| | GH | 365111.3011 | P-value | 0 | 0 | 0 |
| 22-May | Roll | 391784.1099 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 394046.2345 | Chi-stat | 4524.2491 | 6184.3464 | 1660.0972 |
| | GH | 394876.2831 | P-value | 0 | 0 | 0 |
| 23-May | Roll | 504834.4489 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 507961.0369 | Chi-stat | 6253.1761 | 9019.2506 | 2766.0745 |
| | GH | 509344.0742 | P-value | 0 | 0 | 0 |
| 24-May | Roll | 426820.3463 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 429332.9695 | Chi-stat | 5025.2465 | 7617.5436 | 2592.2971 |
| | GH | 430629.1181 | P-value | 0 | 0 | 0 |
| 25-May | Roll | 368283.5937 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 369811.1244 | Chi-stat | 3055.0614 | 4982.7536 | 1927.6922 |
| | GH | 370774.9705 | P-value | 0 | 0 | 0 |
| 26-May | Roll | 409937.9610 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 411698.1906 | Chi-stat | 3520.4591 | 6524.8684 | 3004.4093 |
| | GH | 413200.3952 | P-value | 0 | 0 | 0 |
| 29-May | Roll | 245069.8893 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 245384.3328 | Chi-stat | 628.8869 | 2015.8596 | 1386.9727 |
| | GH | 246077.8191 | P-value | 0 | 0 | 0 |
| 30-May | Roll | 387333.0538 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 388964.1733 | Chi-stat | 3262.2390 | 5358.2592 | 2096.0202 |
| | GH | 390012.1833 | P-value | 0 | 0 | 0 |
| 31-May | Roll | 410784.406 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 412664.8679 | Chi-stat | 3760.9238 | 5602.1671 | 1841.2433 |
| | GH | 413585.4895 | P-value | 0 | 0 | 0 |
| 1-Jun | Roll | 322725.4681 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 324025.2884 | Chi-stat | 2599.6405 | 3893.7146 | 1294.0740 |
| | GH | 324672.3254 | P-value | 0 | 0 | 0 |
| 2-Jun | Roll | 570025.7323 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 571681.6274 | Chi-stat | 3311.7902 | 4866.7338 | 1554.9435 |
| | GH | 572459.0992 | P-value | 0 | 0 | 0 |
| 5-Jun | Roll | 432138.5773 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 434409.7598 | Chi-stat | 4542.3650 | 6720.9511 | 2178.5861 |
| | GH | 435499.0528 | P-value | 0 | 0 | 0 |
| 6-Jun | Roll | 315554.1838 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 316977.6943 | Chi-stat | 2847.0210 | 4228.0255 | 1381.0045 |
| | GH | 317668.1966 | P-value | 0 | 0 | 0 |
| 7-Jun | Roll | 408143.8464 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 410522.4484 | Chi-stat | 4757.2040 | 6898.4437 | 2141.2397 |
| | GH | 411593.0683 | P-value | 0 | 0 | 0 |
| 8-Jun | Roll | 392998.4378 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 394881.2157 | Chi-stat | 3765.5559 | 5779.1010 | 2013.5452 |
| | GH | 395887.9883 | P-value | 0 | 0 | 0 |
| 9-Jun | Roll | 389712.6362 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 391565.2274 | Chi-stat | 3705.1824 | 5361.7774 | 1656.5950 |
| | GH | 392393.5249 | P-value | 0 | 0 | 0 |
| 12-Jun | Roll | 517206.8955 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |

| | | | | | | |
|---|---|---|---|---|---|---|
| | MS | 519606.3378 | Chi-stat | 4798.8845 | 8008.6417 | 3209.7572 |
| | GH | 521211.2164 | P-value | 0 | 0 | 0 |
| 13-Jun | Roll | 489502.5198 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 492309.1285 | Chi-stat | 5613.2175 | 8273.0244 | 2659.8069 |
| | GH | 493639.032 | P-value | 0 | 0 | 0 |
| 14-Jun | Roll | 511532.4208 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 513162.2615 | Chi-stat | 3259.6815 | 5427.8531 | 2168.1716 |
| | GH | 514246.3473 | P-value | 0 | 0 | 0 |
| 15-Jun | Roll | 867118.8519 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 870552.3898 | Chi-stat | 6867.0758 | 12740.2098 | 5873.1340 |
| | GH | 873488.9568 | P-value | 0 | 0 | 0 |
| 16-Jun | Roll | 530716.4036 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 533219.1178 | Chi-stat | 5005.4283 | 7886.3863 | 2880.9580 |
| | GH | 534659.5968 | P-value | 0 | 0 | 0 |
| 19-Jun | Roll | 474641.9255 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 476240.678 | Chi-stat | 3197.5050 | 3469.1268 | 271.6218 |
| | GH | 476376.4889 | P-value | 0 | 0 | 0 |
| 20-Jun | Roll | 519765.9536 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 520691.8947 | Chi-stat | 1851.8821 | 5246.3840 | 3394.5018 |
| | GH | 522389.1456 | P-value | 0 | 0 | 0 |
| 21-Jun | Roll | 369354.6134 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 370201.8094 | Chi-stat | 1694.3920 | 3876.7641 | 2182.3721 |
| | GH | 371292.9955 | P-value | 0 | 0 | 0 |
| 22-Jun | Roll | 444442.935 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 444883.4025 | Chi-stat | 880.9349 | 3413.8757 | 2532.9408 |
| | GH | 446149.8729 | P-value | 0 | 0 | 0 |
| 23-Jun | Roll | 1374533.738 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 1386969.649 | Chi-stat | 24871.8216 | 29075.2245 | 4203.4029 |
| | GH | 1389071.351 | P-value | 0 | 0 | 0 |
| 26-Jun | Roll | 683033.9206 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 684608.0586 | Chi-stat | 3148.2760 | 5604.7667 | 2456.4907 |
| | GH | 685836.304 | P-value | 0 | 0 | 0 |
| 27-Jun | Roll | 480307.2243 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 481121.4758 | Chi-stat | 1628.5031 | 4280.9591 | 2652.4561 |
| | GH | 482447.7039 | P-value | 0 | 0 | 0 |
| 28-Jun | Roll | 449722.4166 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 450602.4876 | Chi-stat | 1760.1420 | 4875.4327 | 3115.2907 |
| | GH | 452160.133 | P-value | 0 | 0 | 0 |
| 29-Jun | Roll | 480259.7110 | Comparison | LR:Roll-MS | LR:Roll-GH | LR:MS-GH |
| | MS | 481791.8537 | Chi-stat | 3064.2856 | 6243.7616 | 3179.4760 |
| | GH | 483381.5918 | P-value | 0 | 0 | 0 |

## Appendix A

### A. Bayesian Estimation of the empirical market microstructure models

We proceed first by discussing briefly Hasbrouck's Gibbs sampling algorithm for the Roll model followed by the estimation algorithms we propose for the case of the extended Roll and Glosten and Harris models with autocorrelated trade direction indicators.

### A.1 Hasbrouck (2004)

To facilitate the explanations to follow, we reproduce the Roll model, equation (1) in the text:

$$p_t = m_t + cq_t$$
$$m_t = m_{t-1} + u_t, \quad u_t \sim N\left(0, \sigma_u^2\right)$$

where $p_t$ is the log transaction price, $m_t$ is the efficient price, $q_t$ is the independent trade direction indicator, and $c$ is the (log) half bid-ask spread. To estimate this model, we use Hasbrouck (2004)'s codes downloaded from his website.

Hasbrouck (2004) draws $c, \sigma_u^2$, and $q\left(\equiv[q_1,...,q_T]\right)$ repeatedly from the posterior conditional density as follows given data $\left(p \equiv [p_1,..., p_T]\right)$:

- Draw $q^{(1)}$ from $f(q_1,...,q_T \mid c^{(0)}, \sigma_u^{(0)}, p)$
- Draw $c^{(1)}$ from $f(c \mid \sigma_u^{(0)}, q^{(1)}, p)$
- Draw $\sigma_u^{(1)}$ from $f(\sigma_u \mid c^{(1)}, q^{(1)}, p)$

The final output is $\left\{c^{(j)}, \sigma_u^{(j)}, q^{(j)}\right\}$ for j = number of simulations.

#### A.1.1 Exact forms of conditional posterior densities

The Roll model can be expressed in regression form as $Y = Xc + u$, $\mathrm{E}(uu') = \sigma_u^2 I_{T-1}$. In the present setup, $Y = [\Delta p_2,...\Delta p_T]$, $X = [\Delta q_2,..., \Delta q_T]$, and $u = [u_2,..., u_T]$. By using conjugate prior distributions, the conditional posterior distributions for the standard regression parameters are given as follows.

**i. Log half spread ($c$):**

- Prior distribution : $c \sim N\left(\mu_c^{prior}, \Omega_c^{prior}\right)$
- Posterior distribution: $c \mid Y, X \sim N\left(\mu_c^{posterior}, \Omega_c^{posterior}\right)$

where $\mu_c^{posterior} = Dd, \Omega_c^{posterior} = D$

$D^{-1} = X'\Omega_u^{-1}X + \left(\Omega_u^{\text{Prior}}\right)^{-1}$ and $d = X'\Omega_u^{-1}Y + \left(\Omega_c^{\text{Prior}}\right)^{-1}\mu_c^{\text{Prior}}$

We impose a non-informative prior for $c$ by assuming a small value for $\mu_c^{\text{Prior}}(10^{-6})$ and large number for $\Omega_c^{\text{Prior}}(10^{6})$. As economic theory dictates the half-spread is positive, Hasbrouck uses a truncated normal density to ensure a $c > 0$.

In this normal linear regression model, the conjugate prior distribution for the variance parameter is the inverted gamma distribution.

**ii. The variance parameter ($\sigma_u^2$):**

- Prior distribution: $\frac{1}{\sigma_u^2} \mid c \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)$
- Posterior distribution: $\frac{1}{\sigma_u^2} \mid c, Y, X \sim \Gamma\left(\frac{\nu_1}{2}, \frac{\delta_1}{2}\right)$

$$\text{where } \nu_1 = \nu_0 + T \quad \delta_1 = \delta_0 + (Y - Xc)'(Y - Xc)$$

**iii. The independent trade direction indicator**

First, we express the joint distribution of $q(\equiv [q_1, ..., q_T])$ using Bayes' theorem and subsequently simplify as in this case $\Pr(q) = 0.5$, and $f(p)$ does not depend on q. For economy of notation, parameters that are given will be dropped from the explicit conditioning set.

$$pr(q \mid p) = f(p \mid q) \times pr(q) \times \frac{1}{f(p)} \propto f(p \mid q)$$

Hasbrouck uses a single-move Gibbs sampling algorithm to draw $q_t$ sequentially as follows.

– Draw $q_1^*$ from $\Pr(q_1 \mid p, q_2, q_3, \ldots, q_T)$
– Draw $q_2^*$ from $\Pr(q_2 \mid p, q_1^*, q_3, \ldots, q_T)$
⋮
– Draw $q_T^*$ from $\Pr(q_T \mid p, q_1^*, q_2^*, \ldots, q_{T-1}^*)$

where $q_t^*$ is the newly drawn value of $q_t$ from those remaining after the previous draw. To complete this algorithm, Hasbrouck derives the conditional distribution $pr(q_t \mid p_t, m_{t-1}, m_{t+1})$ using Bayes' theorem, simplifying it as neither $pr(q_t \mid m_{t-1}, m_{t+1})$ nor $f(p_t \mid m_{t-1}, m_{t+1})$ depend on *q*.

$$p\,\mathrm{r}(q_t \mid p_t, m_{t-1}, m_{t+1}) = \frac{f(p_t \mid q_t, m_{t-1}, m_{t+1}) \times p\,\mathrm{r}(q_t \mid m_{t-1}, m_{t+1})}{f(p_t \mid m_{t-1}, m_{t+1})} \propto f(p_t \mid q_t, m_{t-1}, m_{t+1})$$

Hasbrouck demonstrates that $f(p_t \mid q_t, m_{t-1}, m_{t+1})$ follows a normal distribution $(\phi(\mu_x, \sigma_x, p_t - cq_t) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{((p_t - cq_t) - \mu_x)^2}{2\sigma_x^2}\right))$ and derives posterior densities for three different scenarios. The non-normalized probabilities of a buy and a sell order are determined after evaluating each case at $q_t = +1$ and $q_t = -1$, respectively. The normalized probability of a buy for the three different possibilities is given as follows.

- For t=1,

$$pr(q_1 = 1 \mid p, q_2, q_3, \ldots, q_T) = \frac{\phi(0, \sigma_u, m_2 - p_1 + c)}{\phi(0, \sigma_u, m_2 - p_1 + c) + \phi(0, \sigma_u, m_2 - p_1 - c)},$$

where $m_t = p_t - cq_t$

- For t=2,…..T-1,

$$pr(q_t = 1 \mid p, q_1, ..., q_{t-1}, q_{t+1} \ldots, q_T) = \frac{\phi\left(\frac{1}{2}(m_{t-1} + m_{t+1}), \frac{\sigma_u}{\sqrt{2}}, p_t - c\right)}{\phi\left(\frac{1}{2}(m_{t-1} + m_{t+1}), \frac{\sigma_u}{\sqrt{2}}, p_t - c\right) + \phi\left(\frac{1}{2}(m_{t-1} + m_{t+1}), \frac{\sigma_u}{\sqrt{2}}, c + p_t\right)}$$

- For t=T,

$$pr(q_T = 1 \mid p, q_1, q_2, q_3, \ldots, q_{T-1}) = \frac{\phi(0, \sigma_u, p_T - c - m_{T-1})}{\phi(0, \sigma_u, p_T - c - m_{T-1}) + \phi(0, \sigma_u, p_T + c - m_{T-1})}$$

For each scenario, the Hasbrouck (2004) algorithm draws a random number from the uniform distribution (0,1). If the random number is lower than the normalized probability of a buy it chooses $q_1 = +1$ (and otherwise, $q_1 = -1$).

**A.2 The extended Roll model with an autocorrelated trade direction indicator**

To facilitate the explanations to follow, we first reproduce the extended Roll model in the text:

$$\Delta p_t = c\Delta q_t + u_t, u_t \sim N\left(0, \sigma_u^2\right)$$
where $q_t \in \{1,-1\}, P = pr\left[q_t = 1 \mid q_{t-1} = 1\right]$ and $Q = pr\left[q_t = -1 \mid q_{t-1} = -1\right]$

The above model extends the Roll model (Hasbrouck's algorithm) by incorporating a single move Gibbs sampling algorithm to simulate the autocorrelated and latent trade direction indicator along with its transition probabilities. In sum, the estimation algorithm consists of the following five steps.

- Draw $q^{(1)}$ from $f(q_1,...,q_T \mid c^{(0)}, P^{(0)}, Q^{(0)}, \sigma_u^{(0)}, p)$
- Draw $c^{(1)}$ from $f(c \mid \sigma_u^{(0)}, q^{(1)}, p)$
- Draw $\sigma_u^{(1)}$ from $f(\sigma_u \mid c^{(1)}, q^{(1)}, p)$
- Draw $P^{(1)}$ from $f(P \mid Q^{(0)}, q^{(1)}, p)$
- Draw $Q^{(1)}$ from $f(Q \mid P^{(1)}, q^{(1)}, p)$

The final output is $\left\{c^{(j)}, \sigma_u^{(j)}, P^{(j)}, Q^{(j)}, q^{(j)}\right\}$ for $j$ = number of simulations.

For the half spread ($c$) and the variance parameter **($\sigma_u^2$)**, we still adopt Hasbrouck's algorithms explained in A.1.

**i. The autocorrelated trade direction indicator**

First, after suppressing the conditioning on the parameters, the conditional distribution of $q_t$ is derived via Bayes' rule as follows:

$$pr\left(q_t \mid q_{\neq t}, Y\right) = \frac{f\left(Y \mid q\right) pr\left(q_t \mid q_{\neq t}\right)}{f\left(Y \mid q_{\neq t}\right)} \propto f\left(Y \mid q\right) pr\left(q_t \mid q_{\neq t}\right)$$
where $Y_t = \left(\Delta p_2, \ldots, \Delta p_t\right), Y = Y_T,$ and $q_{\neq t} = \left\{q_1, \ldots, q_{t-1}, q_{t+1}, \ldots q_T\right\},$

The first term in the above equation is:

$$\begin{aligned} f\left(Y \mid q\right) &= f\left(\Delta p_2 \mid \Delta p_1, q\right) \cdots f\left(\Delta p_T \mid Y_{T-1}, q\right) \\ &= f\left(\Delta p_2 \mid q_2, q_1\right) \cdots f\left(\Delta p_t \mid q_t, q_{t-1}\right) \cdots f\left(\Delta p_T \mid q_T, q_{T-1}\right) \\ &\propto f\left(\Delta p_t \mid q_t, q_{t-1}\right) f\left(\Delta p_{t+1} \mid q_{t+1}, q_t\right) \end{aligned}$$

The second line of the above equation arises from the fact that the likelihood function of $\Delta p_t (t = 1,\ldots T)$ depends on the past state variables, $q_{t-1}, q_t$, while the third follows from the fact that all terms except for $f(\Delta p_t \mid q_t, q_{t-1})$ and $f(\Delta p_{t+1} \mid q_{t+1}, q_t)$ are constant, since we are concerned with the probability of $q_t$ conditional on $Y$ and all the other $q_t$'s.

The second term is given by:

$$
\begin{aligned}
pr(q_t \mid q_{\neq t}) &= pr(q_t \mid q_1,\ldots,q_{t-1},q_{t+1},\ldots q_T) \\
&= \frac{pr(q_{t+1},\ldots,q_T \mid q_1,\ldots,q_t)\, pr(q_t \mid q_1,\ldots,q_{t-1})}{pr(q_{t+1,},\ldots,q_T \mid q_1,\ldots,q_{t-1})} \\
&\propto pr(q_{t+1},\ldots,q_T \mid q_1,\ldots,q_t)\, pr(q_t \mid q_1,\ldots,q_{t-1}) \\
&= pr(q_{t+1} \mid q_1,\ldots,q_t)\, pr(q_{t+2} \mid q_1,\ldots,q_{t+1}),\ldots, pr(q_T \mid q_1,\ldots,q_{T-1})\, pr(q_t \mid q_1,\ldots,q_{t-1}) \\
&= pr(q_{t+1} \mid q_t)\, pr(q_{t+2} \mid q_{t+1}),\ldots, pr(q_T \mid q_{T-1})\, pr(q_t \mid q_{t-1}) \\
&\propto pr(q_{t+1} \mid q_t)\, pr(q_t \mid q_{t-1})
\end{aligned}
$$

Applying Bayes' rule to the second line, and the Markov property of state variables to the fifth line and combining, we obtain the final conditional distribution.

$$pr(q_t \mid q_{\neq t}, Y) \propto f(Y \mid q)\, pr(q_t \mid q_{\neq t}) = f(\Delta p_t \mid q_t, q_{t-1})\, f(\Delta p_{t+1} \mid q_{t+1}, q_t)\, pr(q_{t+1} \mid q_t)\, pr(q_t \mid q_{t-1})$$

Let $P_0 = pr(q_t = -1 \mid q_{\neq t}, Y)$ and $P_1 = pr(q_t = 1 \mid q_{\neq t}, Y)$. These are un-normalized probabilities of a sell and a buy order, respectively. By using two values, we obtain the following normalized probabilities:

$$\Pr_0 = \frac{P_0}{P_0 + P_1}$$

If the normalized probability of a sell $(\Pr_0)$ is higher (lower) than a value from the uniform distribution (0,1), we set $q_t = -1$ (+1). Specifically, the simulation algorithm can be stated as follows.

- For t = 0,

We use the unconditional probability to determine the direction of $q_0$

$$pr[q_0 = 1] = \frac{1-Q}{2-P-Q}$$

If $pr[q_0 = 1]$ is higher than a draw from uniform distribution [0,1], then $q_0 = 1$.

- For t = 1,…..T-1,

$$pr\left(q_t \mid q_{\neq t}, Y\right) \propto f\left(\Delta p_t \mid q_t, q_{t-1}\right) f\left(\Delta p_{t+1} \mid q_{t+1}, q_t\right) pr\left(q_{t+1} \mid q_t\right) pr\left(q_t \mid q_{t-1}\right)$$

where $pr\left(q_t = -1 \mid q_{\neq t}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\left(\Delta p_t - c(-1-q_{t-1})\right)^2 + \left(\Delta p_{t+1} - c\left(q_{t+1}+1\right)\right)^2\right)\right\}$

$$\cdot pr\left(q_{t+1} \mid q_t = -1\right) \cdot pr\left(q_t = -1 \mid q_{t-1}\right)$$

$$pr\left(q_t = 1 \mid q_{\neq t}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma^2}\left(\left(\Delta p_t - c(1-q_{t-1})\right)^2 + \left(\Delta p_{t+1} - c\left(q_{t+1}-1\right)\right)^2\right)\right\}$$

$$\cdot pr\left(q_{t+1} \mid q_t = 1\right) \cdot pr\left(q_t = 1 \mid q_{t-1}\right)$$

- For t = T,

$$pr\left(q_T \mid q_{\neq T}, Y\right) \propto f\left(\Delta p_T \mid q_T, q_{T-1}\right) pr\left(q_T \mid q_{T-1}\right)$$

where $pr\left(q_T = -1 \mid q_{\neq T}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\left(\Delta p_T - c(-1-q_{T-1})\right)^2\right\} pr\left(q_T = -1 \mid q_{T-1}\right)\right.$

$$pr\left(q_T = 1 \mid q_{\neq T}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\left(\Delta p_T - c(1-q_{T-1})\right)^2\right\} pr\left(q_T = 1 \mid q_{T-1}\right)\right.$$

**ii. Transition probabilities:**

Finally, we are required to estimate the two transition probabilities *P* and *Q*. Following the discussion contained in Chapter 9 of Kim and Nelson (1999), we use the beta distribution as follows:

- Independent prior distribution :

$$P \sim beta\left(u_{11}, u_{1-1}\right) = P^{u_{11}-1}\left(1-P\right)^{u_{1-1}-1}$$

$$Q \sim beta(u_{-1-1}, u_{1-1}) = Q^{u_{-1-1}-1}\left(1-Q\right)^{u_{-11}-1}$$

where $u_{ij}\left(i, j = -1,1\right)$ are known parameters of the priors.

- Posterior distribution :

$$P \mid q \sim beta\left(u_{11} + n_{11}, u_{1-1} + n_{1-1}\right)$$

$$Q \mid q \sim beta(u_{-1-1} + n_{-1-1}, u_{-11} + n_{-11})$$

where $n_{ij}$ refers to the transitions from state $i$ to $j$, which can be easily counted for given $q$.

**A3. The extended Glosten and Harris Model**

To facilitate the explanations to follow, we reproduce the extended Glosten and Harris model in the text:

$$\Delta p_t = c\Delta q_t + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \left((P-Q)+(P+Q-1)q_{t-1}\right)\right) + u_t, \quad u_t \sim N\left(0,\sigma_u^2\right)$$

In this model, we have two extra parameters ( $\lambda_0$ and $\lambda_1$ ) and two transition probabilities ( $P$ and $Q$ ) appearing in the main regression. Therefore, the conditional posterior densities for the transition probabilities parameters do not follow any known distributions. Therefore the Gibbs sampling estimation algorithm as employed in the previous section is not feasible to estimate these parameters. To overcome this issue, we adopt a variant of the tailored Random-Walk Metropolis-Hastings Algorithm developed by Chib and Ramamurthy (2010).

Specifically, we draw $c, \lambda_0, \lambda_1, P, Q, \sigma_u$, and $q$ repeatedly as follows. First, the parameters and latent trade direction indicators are set equal to some arbitrary values. Denoting these initial values as $\{\left[c^{(0)}, \lambda_0^{(0)}, \lambda_1^{(0)}, P^{(0)}, Q^{(0)}\right], \sigma_u^{(0)}, q^{(0)}\}$ , the first step is represented as follows.

- Draw $q^{(1)}$ from $f(q|\left[c^{(0)}, \lambda_0^{(0)}, \lambda_1^{(0)}, P^{(0)}, Q^{(0)}\right], \sigma_u^{(0)}, Y)$
- Draw $\left[c^{(1)}, \lambda_0^{(1)}, \lambda_1^{(1)}, P^{(1)}, Q^{(1)}\right]$ from $f(c, \lambda_0, \lambda_1, P, Q | \sigma_u^{(0)}, q^{(1)}, Y)$
- Draw $\sigma_u^{(1)}$ from $f(\sigma_u | \left[c^{(1)}, \lambda_0^{(1)}, \lambda_1^{(1)}, P^{(1)}, Q^{(1)}\right], q^{(1)}, Y)$

By repeating this procedure many times, we generate a sequence of draws of unknowns for $j = 1, \ldots, n$. The Gibbs principle demonstrates that the limiting distribution of the $n^{th}$ draw after burn-in samples (as $n \to \infty$) is $f(\left[c, \lambda_0, \lambda_1, P, Q\right], \sigma_u, q | Y)$, the desired posterior, and the limiting draw for any parameter is distributed as the corresponding marginal posterior. The details for Bayesian algorithms are presented below.

**i. The autocorrelated trade direction indicator**

Once we suppress the conditioning on the parameters, the conditional distribution of $q_t$ is derived via Bayes' rule following the procedure outlined in A.2.

$$pr\left(q_t \mid q_{\neq t}, Y\right) \propto f\left(Y \mid q\right) pr\left(q_t \mid q_{\neq t}\right) \propto f\left(\Delta p_t \mid q_t, q_{t-1}\right) f\left(\Delta p_{t+1} \mid q_{t+1}, q_t\right) pr\left(q_{t+1} \mid q_t\right) pr\left(q_t \mid q_{t-1}\right)$$

Define $P_0 = pr\left(q_t = -1 \mid q_{\neq t}, Y\right)$ and $P_1 = pr\left(q_t = 1 \mid q_{\neq t}, Y\right)$ as the non-normalized probabilities of a sell and a buy trade. By using these two values, we obtain the following normalized probabilities for $q_t = -1$.

$$\Pr_0 = \frac{P_0}{P_0 + P_1}$$

If the normalized probability of a sell $\left(\Pr_0\right)$ is higher (lower) than a value from the uniform distribution (0,1), we set $q_t = -1$ (+1). Specifically, the simulation algorithm can be stated as follows.

Specifically, the simulation algorithm can be stated as follows.

- For t = 0,

  We use the unconditional probability to determine the $q_0$.

  $$pr\left[q_0 = 1\right] = \frac{1-Q}{2-P-Q}$$

  If $pr\left[q_0 = 1\right]$ is higher than a draw from uniform distribution [0,1], then $q_0 = 1$.

- For t = 1,…..T-1,

  $$pr\left(q_t \mid q_{\neq t}, Y\right) \propto f\left(\Delta p_t \mid q_t, q_{t-1}\right) f\left(\Delta p_{t+1} \mid q_{t+1}, q_t\right) pr\left(q_{t+1} \mid q_t\right) pr\left(q_t \mid q_{t-1}\right)$$

  $$\text{where } pr\left(q_t = -1 \mid q_{\neq t}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\begin{array}{l}\left(\left(\Delta p_t - c(-1-q_{t-1}) + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right) \times \left(1 + \mu + \rho q_{t-1}\right)\right)^2\right. \\ \left. + \left(\Delta p_{t+1} - c\left(q_{t+1}+1\right) - \left(\lambda_0 + \lambda_1\sqrt{V_{t+1}}\right) \times \left(q_{t+1} - \mu + \rho\right)\right)^2\end{array}\right)\right\}$$

  $$\bullet\, pr\left(q_{t+1} \mid q_t = -1\right) \times pr\left(q_t = -1 \mid q_{t-1}\right)$$

  $$pr\left(q_t = 1 \mid q_{\neq t}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\begin{array}{l}\left(\left(\Delta p_t - c(1-q_{t-1}) + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right) \times \left(-1 + \mu + \rho q_{t-1}\right)\right)^2\right. \\ \left. + \left(\Delta p_{t+1} - c\left(q_{t+1}-1\right) - \left(\lambda_0 + \lambda_1\sqrt{V_{t+1}}\right) \times \left(q_{t+1} - \mu - \rho\right)\right)^2\end{array}\right)\right\}$$

  $$\bullet\, pr\left(q_{t+1} \mid q_t = 1\right) \times pr\left(q_t = 1 \mid q_{t-1}\right)$$

where $\mu = P - Q$ and $\rho = P + Q - 1$

- For t = T,

$$pr\left(q_T \mid q_{\neq T}, Y\right) \propto pr\left(\Delta p_T \mid q_T, q_{T-1}\right) pr\left(q_T \mid q_{T-1}\right)$$

where $pr\left(q_T = -1 \mid q_{\neq T}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\left(\Delta p_T - c(-1 - q_{T-1}) + \left(\lambda_0 + \lambda_1\sqrt{V_T}\right) \times \left(1 + \mu + \rho q_{T-1}\right)\right)^2\right\}\right.$

$$\bullet p\left(q_T = -1 \mid q_{T-1}\right)$$

$$P_1 = p\left(q_T = 1 \mid q_{\neq T}, Y\right) \propto \exp\left\{-\frac{1}{2\sigma_u^2}\left(\left(\Delta p_T - c(1 - q_{T-1}) + \left(\lambda_0 + \lambda_1\sqrt{V_T}\right) \times \left(-1 + \mu + \rho q_{T-1}\right)\right)^2\right\}\right.$$

$$\bullet p\left(q_T = 1 \mid q_{T-1}\right)$$

**ii. The mean parameters** $\left(c, \lambda_0, \lambda_1, P, Q\right)$

We use the tailored random walk Metropolis Hastings algorithm to draw $\theta = \left(c, \lambda_0, \lambda_1, P, Q\right)$ given other model parameters and the latent variables ($q$) with the positivity restriction on ($c$>0)

In order to compute an acceptance/rejection probability for the Metropolis Hastings algorithm, we need to derive the joint conditional posterior density of $\theta = \left(c, \lambda_0, \lambda_1, P, Q\right)$.

- Prior distribution of $\theta = \left(c, \lambda_0, \lambda_1, P, Q\right)$**:**

We assume independent prior distributions for the following parameters.

a. Multivariate normal density prior for $\beta = \left(c, \lambda_0, \lambda_1\right)$

: $\beta \sim N\left(\beta_0, \Sigma_0\right) = \left(2\pi\right)^{-\frac{3}{2}} \left|\Sigma_0\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\beta - \beta_0)' \Sigma_0^{-1} (\beta - \beta_0)}$

b. Beta density prior for $P$ and $Q$

: $beta\left(P, Q\right) = P^{u_{1,1}-1}\left(1 - P\right)^{u_{1,-1}-1} Q^{u_{-1,-1}-1}\left(1 - Q\right)^{u_{-1,1}-1}$

Therefore, the joint prior density for $\theta = \left(c, \lambda_0, \lambda_1, P, Q\right)$ is as follows.

$$P^{u_{11}-1}Q^{u_{-1-1}-1}(1-Q)^{u_{-11}-1}(1-P)^{u_{1-1}-1}\times(2\pi)^{-\frac{3}{2}}|\Sigma_0|^{-\frac{1}{2}}e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)}$$

- Conditional likelihood function given $q$

The conditional likelihood function $L(c,\lambda_0,\lambda_1,P,Q\,|\,q,Y)$ of the extended Glosten and Harris (1988) model is formed by suppressing the conditioning on all the parameters and data except for $q$ and $Y$ as follows.

$$L(c,\lambda_0,\lambda_1,P,Q\,|\,q,Y)=f(Y\,|\,q)\,pr(q)$$

Where $f(Y\,|\,q)=f(y_1\,|\,q_1,q_0,\tilde{y}_0)\,f(y_2\,|\,q_2,q_1,\tilde{y}_1)\cdots f(y_T\,|\,q_T,q_{T-1},\tilde{y}_{T-1})$

$f(q)=f(q_1\,|\,q_0)\,f(q_2\,|\,q_1)\cdots f(q_T\,|\,q_{T-1})$ and $\tilde{y}_{t-1}=\{y_1,y_2,\cdots,y_{t-1}\}$

We can express the above likelihood function in the following compact form.

$$L(c,\lambda_0,\lambda_1,P,Q\,|\,q,Y)=f_{ij}(1)\times\cdots\times f_{ij}(T)\times P^{n_{11}}Q^{n_{-1-1}}(1-Q)^{n_{-11}}(1-P)^{n_{1-1}}$$

where $i,j=\{1,-1\}$ and $n_{i,j}$ is the total number of transitions from state $q_{t-1}=i$ to $q_t=j$, for t=1, 2,..,T. And $f_{ij}(t)=f(\Delta p_t\,|\,q_t=j,q_{t-1}=i,\tilde{y}_{t-1})$

$$=\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left\{-\frac{1}{2\sigma_u^2}\left(\Delta p_t-c\left(q_t^j-q_{t-1}^i\right)-\left(\lambda_0+\lambda_1\sqrt{V_t}\right)\left(q_t^j-\left(\mu+\rho q_{t-1}^i\right)\right)\right)^2\right\}$$

- Conditional Posterior density of $\theta=(c,\lambda_0,\lambda_1,P,Q)$ given $q$

By multiplying the joint prior density by the conditional likelihood function, we obtain the following conditional posterior density $f(c,\lambda_0,\lambda_1,P,Q\,|\,q,Y)$.

$$f(c,\lambda_0,\lambda_1,P,Q\,|\,\tilde{q}_T)\propto f_{ij}(1)\times\cdots\times f_{ij}(T)\times P^{u_{11}+n_{11}-1}Q^{u_{-1-1}+n_{-1-1}-1}(1-Q)^{u_{-11}+n_{-11}-1}(1-P)^{u_{1-1}+n_{1-1}-1}$$
$$\times(2\pi)^{-\frac{3}{2}}|\Sigma_0|^{-\frac{1}{2}}e^{-\frac{1}{2}(\beta-\beta_0)'\Sigma_0^{-1}(\beta-\beta_0)}$$

- Joint proposal density $\theta=(c,\lambda_0,\lambda_1,P,Q)$

We follow the procedure underpinning the joint proposal density of the tailored Random-Walk Metropolis-Hastings Algorithm developed by Chib and Ramamurthy (2010) as follows.

a. We first maximize the log conditional posterior density using a simulated annealing algorithm or another robust maximization algorithm (e.g. fminsearch in Matlab) to obtain maximum posterior estimates $\theta^* = \left(c^*, \lambda_0^*, \lambda_1^*, P^*, Q^*\right)$ and corresponding variance estimate $Var\left(\theta^*\right)$. Denote the old (new) parameter value $\theta^{old}\left(\theta^{new}\right)$ in the MCMC iteration. [21]

b. We simulate a new candidate $\theta^{new}$ using the following proposal density, (denoted as $q^*\left(\theta^{new} \mid .\right)$).

Multivariate t -density ( $q^*\left(\theta^{new} \mid \theta^*, Var\left(\theta^*\right), nu\right)$).

$$\theta^{new} = MVT\left(\theta^*, Var\left(\theta^*\right), nu\right)$$

where $nu$ is the degree of freedom parameter, and will be set to obtain 20%-50% of acceptance ratio for $\theta^{new}$ defined below.

Note: if $\theta^{new}$ satisfies the several restrictions on $\theta$ (e.g. non-negativity restrictions for $c, P, Q$) then we proceed to the next step. If that is not the case, we set $\theta = \theta^{old}$ and terminate the draw at this point.

c. We compute the following ratio to accept or reject the proposed value for $\theta^{new}$.

$$\alpha_\theta = \frac{f\left(\theta^{new} \mid \tilde{q}_T\right) \times q^*\left(\theta^{old} \mid .\right)}{f\left(\theta^{old} \mid \tilde{q}_T\right) \times q^*\left(\theta^{new} \mid .\right)}$$

To implement the accept-reject step, we draw a uniform random variable, $U \sim U[0,1]$, and set $\theta = \theta^{new}$ if $U < \alpha_\theta$ and $\theta = \theta^{old}$ if $U > \alpha_\theta$.

[21] In the new Tailored MH algorithms, in each step of MCMC, we estimate the mode and variance of the proposal densities by maximizing the log posterior densities of all parameters using simulated annealing or Nelder-Mead simplex algorithm.

**iv. The variance parameter ($\sigma_u^2$):**

We need to express first the Glosten and Harris (1988) model in the following regression format.

$$\Delta p_t = \left[ q_t - q_{t-1}, -\mu + q_t - \rho q_{t-1}, q_t\sqrt{V_t} - \mu\sqrt{V_t} - \rho q_{t-1}\sqrt{V_t} \right] \begin{bmatrix} c \\ \lambda_0 \\ \lambda_1 \end{bmatrix} + u_t$$

In matrix notation,

$$y_t = X_t\beta + u_t, u_t \sim N\left(0, \sigma_u^2\right)$$

$$\text{where } y_t = \Delta p_t, X_t = \begin{bmatrix} q_t - q_{t-1} \\ -\mu + q_t - \rho q_{t-1} \\ q_t\sqrt{V_t} - \mu\sqrt{V_t} - \rho q_{t-1}\sqrt{V_t} \end{bmatrix}', \beta = \begin{bmatrix} c \\ \lambda_0 \\ \lambda_1 \end{bmatrix}$$

Based on this notation, we can express the conditional posterior distribution for $\sigma_u^2$ as follows.

- Prior distribution: $\frac{1}{\sigma_u^2} | \theta \sim \Gamma\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)$
- Posterior distribution: $\frac{1}{\sigma_u^2} | \theta, Y, X \sim \Gamma\left(\frac{\nu_1}{2}, \frac{\delta_1}{2}\right)$

$$\text{where } \nu_1 = \nu_0 + T, \delta_1 = \delta_0 + (Y - X\beta)'(Y - X\beta)$$

## Appendix B

### B. MCMC convergence diagnostics

In any MCMC estimation, it is crucial to determine the convergence of the chain to conduct a correct statistical inference. In this paper, as a diagnostic to check the convergence of our MCMC algorithm, we compute and report the effective sample size based on the inefficiency factors and the p-value of Geweke's convergence diagnostics test for the model parameters. Specifically, we start all estimations with 100,000 burn-in period and 250,000 total numbers of iterations and increase these numbers by 10,000 until the convergence criteria for both (explained below) are satisfied simultaneously for all parameters.

#### B.1 Geweke's convergence diagnostic (CD)

The idea of Geweke's diagnostic is simple and mimics the two-sample test of means. First we set the total number of iterations ($n$), the burn-in period ($n_0$), and the rest of iterations ($n_1$) (i.e., $n = n_0 + n_1$). Then divide $n_1$ into three periods (e.g., $n_1 = n_A + n_B + n_C$). Specifically, we set the first 40% ($n_A$), the second 20% ($n_B$), and the last 40% ($n_C$) as three sub-periods. And if the mean of the first 40% is not significantly different from that of the last 40%, then we conclude the target distribution converged somewhere in the first 40% of the chain.

More formally, we compute Geweke's CD as follows.

$$CD = \left(\hat{\theta}_{n_A} - \hat{\theta}_{n_C}\right) / \left(\frac{\hat{\sigma}_A}{\sqrt{n_A}} + \frac{\hat{\sigma}_C}{\sqrt{n_C}}\right)$$

where $\hat{\theta}_{n_A}, \hat{\theta}_{n_B}, \hat{\theta}_{n_C}$ ($\hat{\sigma}_A, \hat{\sigma}_B, \hat{\sigma}_C$) are the sample means (standard deviation) of each sub-period. Based on the Geweke's CD, we compute Geweke's p-value at 0.05 significance level to test if the means of the first and the last samples are same.

$$Geweke'\ p = 2 \times \left[1 - \Phi\left(|CD|\right)\right]$$

where $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution. If Geweke's p-value is less than 0.05, we interpret the burn-in period is too small and we can't

guarantee the convergence of the chain. In general, as Geweke's p-value is close to 1, we have more efficient samples.

### B.2. Inefficiency factor and effective sample size

The inefficiency factor, as proposed by Kim, Shephard, Chib (1998), is defined as $1+2\sum_{k=1}^{\infty}\rho(k)$ where $\rho(k)$ is the $k^{th}$ order autocorrelation coefficient and measures how well the MCMC sequence mixes. In this paper, we estimate it as $1+(2\times 200)/(200-1)\sum_{j=1}^{200}K(j/\mathrm{B})\hat{\rho}(j)$ where $\hat{\rho}(j)$ is the $j^{th}$ order sample autocorrelation coefficient of the MCMC draws and $K(\cdot)$ stands for the Parzen kernel. A value of 1 indicates that the MCMC draws are uncorrelated with a good mixing, while large values indicate a slow mixing. The effective sample size is computed as the ratio of the number of iterations after the burn-in period to the inefficiency factor and should be larger than 1000 for all parameters to guarantee sufficient number of the MCMC draws.

## Appendix C

### C. Maximum Likelihood Estimation (MLE) methods of the empirical market microstructure models

The empirical market microstructure models employed in this paper imply that we can interpret them as a regime switching model. Based on the Hamilton filter, we obtain a likelihood function and subsequently use MLE methods for estimation (see chapter 4 of Kim and Nelson (1999) for more details).

As an illustration, we demonstrate how to construct the likelihood function for the Glosten and Harris model with autocorrelated $q_t$. This model implies that:

$$\Delta p_t = c\Delta q_t + \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \left((P-Q)+(P+Q-1)q_{t-1}\right)\right) + u_t, \quad u_t \square N\left(0,\sigma_u^2\right)$$

where $p_t$ is the log transaction price, $q_t$ is the trade direction indicator with transition probabilities $P = pr(q_t = 1 \mid q_{t-1} = 1)$ and $Q = pr(q_t = -1 \mid q_{t-1} = -1)$.

On this basis, we interpret this form of the Glosten and Harris model with autocorrelated $q_t$ as a standard regime switching model. The likelihood function $L\left(c,\lambda_0,\lambda_1,P,Q \mid \Delta p_2,\ldots,\Delta p_T\right)$ of the extended Glosten and Harris (1988) model is formed by suppressing the conditioning on all the parameters.

$$: L\left(c,\lambda_0,\lambda_1,P,Q \mid \Delta p_2,\ldots,\Delta p_T\right) = f\left(\Delta p_3 \mid \Delta p_2\right)\cdots f\left(\Delta p_T \mid \Delta \tilde{p}_{T-1}\right)$$

where $\Delta \tilde{p}_{t-1} = \left\{\Delta p_2,\cdots,\Delta p_{t-1}\right\}$ and

$$\begin{aligned} f\left(\Delta p_t \mid \Delta \tilde{p}_{t-1}\right) &= \sum_{q_t}\sum_{q_{t-1}} f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta \tilde{p}_{t-1}\right) pr\left(q_t, q_{t-1} \mid \Delta \tilde{p}_{t-1}\right) \\ &= \sum_{q_t}\sum_{q_{t-1}} f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta \tilde{p}_{t-1}\right) pr\left(q_t \mid q_{t-1}\right) pr\left(q_{t-1} \mid \Delta \tilde{p}_{t-1}\right) \end{aligned}$$

First the conditional likelihood function $f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta \tilde{p}_{t-1}\right)$ is expressed as follows.

$$f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta\tilde{p}_{t-1}\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{1}{2\sigma_u^2}\left(\Delta p_t - c\Delta q_t - \left(\lambda_0 + \lambda_1\sqrt{V_t}\right)\left(q_t - \left(\left(P-Q\right) + (P+Q-1)q_{t-1}\right)\right)\right)^2\right\}$$

Second, in order to complete this likelihood function, we need to compute

$pr(q_t \mid \Delta\tilde{p}_{t-1}) = \sum_{q_{t-1}} pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1})$ based on the Hamilton filter. The Hamilton filter for this model consists of the following:

$$pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1}) = pr(q_t \mid q_{t-1})\, pr(q_{t-1} \mid \Delta\tilde{p}_{t-1})$$

We are required to initialize $P(q_1 \mid \Delta\tilde{p}_0)$ using the steady-state probabilities. In order to get $P(q_t \mid \Delta\tilde{p}_t)$ for the next iteration, we need to compute the following equations repeatedly for all time t:

$$pr(q_t, q_{t-1} \mid \Delta\tilde{p}_t) = \frac{f(\Delta p_t, q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1})}{f(\Delta p_t \mid \Delta\tilde{p}_{t-1})} = \frac{f(\Delta p_t \mid q_t, q_{t-1}\Delta\tilde{p}_{t-1})\, pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1})}{f(\Delta p_t \mid \Delta\tilde{p}_{t-1})}$$ and

$$pr(q_t \mid \Delta\tilde{p}_t) = \sum_{q_{t-1}} pr(q_t, q_{t-1} \mid \Delta\tilde{p}_t)$$

where $f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta\tilde{p}_{t-1}\right)$ is already given in the above. In sum, an MLE for this model can be developed by summing $f\left(\Delta p_t \mid \Delta\tilde{p}_{t-1}\right)$ using the probability terms ($pr\left(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1}\right)$) we computed in the Hamilton filter for each regime over the whole sample.

$$L\left(c, \lambda_0, \lambda_1, P, Q \mid \Delta\tilde{p}_t\right) = \sum_{t=2}^{T} \ln\left(f\left(\Delta p_t \mid \Delta\tilde{p}_{t-1}\right)\right) = \sum_{t=2}^{T} \ln\left(\sum_{q_t}\sum_{q_{t-1}} f(\Delta p_t \mid q_t, q_{t-1}, \Delta\tilde{p}_{t-1})\, pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1})\right)$$

As nested special cases of this model, we can construct the likelihood functions of the Roll and the extended Roll model with autocorrelated $q_t$ as follows.

Case 1) the Roll model with $\lambda_0 = \lambda_1 = 0$ and $P = Q = 1/2$

Where $f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta\tilde{p}_{t-1}\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{1}{2\sigma_u^2}\left(\Delta p_t - c\Delta q_t\right)^2\right\}$ and

$pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1}) = 1/2\, pr(q_{t-1} \mid \Delta\tilde{p}_{t-1})$

Case 2) the extended Roll model with $\lambda_0 = \lambda_1 = 0$

where $f\left(\Delta p_t \mid q_t, q_{t-1}, \Delta\tilde{p}_{t-1}\right) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{1}{2\sigma_u^2}\left(\Delta p_t - c\Delta q_t\right)^2\right\}$ and

$pr(q_t, q_{t-1} \mid \Delta\tilde{p}_{t-1}) = pr(q_t \mid q_{t-1})\, pr(q_{t-1} \mid \Delta\tilde{p}_{t-1})$

## Appendix D

### D. Kim (1994)'s smoothing algorithm.

Given MLE estimates of the model, we can make inferences on the trade direction indicator $q_t$ using all the information in the sample (called smoothing). This gives us the probability of buy-initiated (sell-initiated) trade $\Pr[q_t = 1 | \Delta\tilde{p}_T]$ ( $\Pr[q_t = -1 | \Delta\tilde{p}_T]$ ) for every transaction.

When the models in this paper are estimated by MLE as shown in appendix C, we already utilized Hamilton filter to construct the likelihood function. The Hamilton filter uses the information available up to time *t* to compute the buy-initiated (sell-initiated) trade at time t, $\Pr[q_t = 1 | \Delta\tilde{p}_t]$ ($\Pr[q_t = -1 | \Delta\tilde{p}_t]$).

Consider the following derivation of the joint probability that $q_t = j$ and $q_{t+1} = k$ based on full information:

$$\Pr[q_t = j,\ q_{t+1} = k | \Delta\tilde{p}_T]$$

$$= \Pr[q_{t+1} = k | \Delta\tilde{p}_T] \times \Pr[q_t = j | q_{t+1} = k, \Delta\tilde{p}_T]$$

Chapter 4 of Kim and Nelson (1999) explains that $\Pr[q_t = j | q_{t+1} = k, \Delta\tilde{p}_T]$ $= \Pr[q_t = j | q_{t+1} = k, \Delta\tilde{p}_t]$ because if $q_{t+1}$ were somehow known, then $(\Delta p_{t+1}, \Delta p_{t+2}, ..., \Delta p_T)$ would contain no information about $q_t$ beyond that contained in $q_{t+1}$ and $\Delta\tilde{p}_t$ .

$$= \Pr[q_{t+1} = k | \Delta\tilde{p}_T] \times \Pr[q_t = j | q_{t+1} = k, \Delta\tilde{p}_t]$$

$$= \frac{\Pr[q_{t+1} = k | \Delta\tilde{p}_T] \times \Pr[q_t = j, q_{t+1} = k | \Delta\tilde{p}_t]}{\Pr[q_{t+1} = k | \Delta\tilde{p}_t]}$$

$$= \frac{\Pr[q_{t+1} = k | \Delta\tilde{p}_T] \times \Pr[q_t = j | \Delta\tilde{p}_t] \times \Pr[q_{t+1} = k | q_t = j]}{\Pr[q_{t+1} = k | \Delta\tilde{p}_t]}$$

and

$$\Pr[q_t = j | \Delta\tilde{p}_T] = \Pr[q_t = j, q_{t+1} = -1 | \Delta\tilde{p}_T] + \Pr[q_t = j, q_{t+1} = 1 | \Delta\tilde{p}_T] \text{ for } j = -1 \text{ and } 1.$$

Given $\Pr[q_T \mid \Delta\tilde{p}_T]$ at the last iteration of the Hamilton filter, the above can be iterated for $t = T-1, T-2, \ldots, 2$ to get the smoothed probabilities, $\Pr[q_t \mid \Delta\tilde{p}_T], t = T-1, T-2, \ldots, 2$

Once we obtain $\Pr[q_t = 1 \mid \Delta\tilde{p}_T]$ and $\Pr[q_t = -1 \mid \Delta\tilde{p}_T]$, we can divide the sample into the two regimes using the following rule:

If $\Pr[q_t = 1 \mid \Delta\tilde{p}_T] > 0.5$, we set $q_t = 1$

If $\Pr[q_t = -1 \mid \Delta\tilde{p}_T] > 0.5$, we set $q_t = -1$

The above rule is the one we propose to classify the trades. As an illustration purpose, we plot the estimated trade direction indicator series of the extended GH model using the first 1000 observations on May 1.