Probability of Informed Trading and Volatility for an ETF

Dimitrios Karyampas*

Credit Suisse AG

Paola Paiardini†

School of Economics and Finance, Queen Mary, University of London

Preliminary and Incomplete

Abstract

We use the new procedure developed by Easley et al. (2010b)to estimate the Probability of Informed Trading (PIN), based on the volume imbalance: Volume-Synchronized Probability of Informed Trading (VPIN). Unlike the previous method, this one does not require the use of numerical methods to estimate unobservable parameters. We also relate the VPIN metric to volatility measures. However, we use most efficient estimators of volatility which consider the number of jumps. Moreover, we add the VPIN to a Heterogeneous Autoregressive model of Realized Volatility to further investigate its relation with volatility. For the empirical analysis we use data on an exchange traded fund (SPY).

JEL codes: C22, C53, D53, G10, G14

Keywords: Market Microstructure, Probability of Informed Trading, VPIN, Jumps

*Credit Suisse AG. *e-mail*: d.karyampas@ems.bbk.ac.uk.
This work expresses the opinion of the author and is in no way representing the opinion of the institution the author works for.

†Queen Mary, University of London, Mile End Road, London, E1 4NS.
e-mail: p.paiardini@qmul.ac.uk

1 Introduction

It is common knowledge in the microstructure literature that the order arrivals contain important information to determine subsequents price movements. Over the years, there have been developed several methods to extract this information from order flow. However, things become more difficult when we need to measure the order flow in a high frequency scenario.

In a framework where trading takes place in milliseconds, trading time loses its meaning. Easley and O'Hara (1992) introduce the concept of time instead of clock time, following the idea that trading time elapsed between two trades gather information. Indeed, this a no-trade interval can occur both when there has not been any information event and when a trader decides not to trade for portfolio reasons.

Easley et al. (2010b) introduce another concept of trading time, as captured by volume. Their idea is that the more relevant a piece of information is, the more volume it will move. Moreover, time ranges do not contain comparable amounts of information. Instead volume buckets are comparable. They suggest to sample the data on the base of volume, so every time that market exchanges a constant amount of volume means that news of comparable relevance are arrived to the market. In this way they compute the so-called *Volume-Synchronized Probability of Informed Trading* (VPIN).

The advantage of this new procedure is that since sampling by volume is a proxy for sampling by volatility (because large price movements are associated with large volumes), it is possible to reduce the impact of volatility clustering in the data. Thus, sampling by volume instead of time allows us to have a dataset which is less heteroskedastic and follows a distribution that is closed to normal.

Moreover, the approach based on the VPIN does not require the numerical maximum likelihood estimation of unobservable parameters, so it allows to overcome all the difficulties of estimating the Probability of Informed Trading (PIN) with the previous technique.

In this paper we apply this new technique to the exchange traded fund tracking the S&P 500 index; the S&P 500 SPDR traded on Amex with ticker SPY. Furthermore, we investigate the possible links between VPIN and volatility, adding the number of jumps and the VPIN to a Heterogenous Autoregressive model of Realized Volatility (HAR-RV).

The paper is organized as follows. Section 2 summarizes the principal approaches to analyse the effects of asymmetric information among market participants on the outcomes of the trading process. Section 3 illustrates the principal characteristics of a benchmark sequential trade model. Section 4 briefly illustrates the VPIN model. Section 5 discusses the different measures of volatility that we use. Section 6 describes the procedure to detect jumps and the measure of volatility that suits better in presence of jumps. Section 7 describes the data and the institutional environment. Section 8 summarizes the empirical results. Section 9 concludes.

2 Asymmetric Information and the Probability of Informed Trading

O'Hara (1998) distinguish between two approaches to analyse the effects of asymmetric information among market participants on the outcomes of the trading process: Walrasian batch (or strategic trader) models and sequential trade models.

In the Walrasian approach, market makers observe the net order flow from traders and set a single price, at which all orders are being executed. The trading process is outlined as a sequence of auctions based on requests to buy or sell a specified number of securities at the market price. These models do not allow to characterize the bid-ask spread, but focus on the effects of order placement by informed and uniformed traders on prices. The pivotal work in this approach is the paper by Kyle (1985).

The sequential trade approach to modelling the security market behaviour in the presence of asymmetric information has been widely used. The trading process is characheterized by traders arrive randomly, singly, sequentially and independently. In these models the order flow is informative because at each round trading behaviour can reveal private information and affect prices. The market maker is risk neutral and sets quotes to maximize his expected profit. He cannot identify informed traders, however he knows the probability that any trade comes from an informed or from an uninformed trader. Thus, the market maker sets quotes taking into account the probabilities of losing to the informed and gaining to uninformed.

The learning process from the order flow has been analysed in a dynamic framework by Glosten and Milgrom (1985) and Easley and O'Hara (1987).Informed traders will reveal their information by selling the asset, if they know bad news and by buying it if they observe good news. Therefore, the fact that someone wants to sell may be interpreted as a signal to the market maker that bad news have been received by this trader, but it may also mean that the trader is uninformed and simply needs liquidity. In Glosten and Milgrom (1985) the market maker adjusts his beliefs about the value of the asset, through a Bayesian learning process on the type of trades he observes.

Easley and O'Hara (1987) introduce the notion of information uncertainty. Glosten and Milgrom (1985) assume that even if the market maker does not observe the information event, he knows that there are traders who always know either good or bad news. In this work, alike Glosten and Milgrom (1985), it is introduced a third possibility, the absence of any private information. In this case the market maker will receive orders only from uninformed traders.

Easley et al. (1996) set up a mixed discrete and continuous time sequential trade model of the trading process, that can be used as a generic framework to analyse the other extensions. We will briefly explain it in the next paragraph. Other versions add trade size and time effects to the basic sequential trade model.

In Easley et al. (1996) no-trade intervals may arise in the course of the trading day but they are treated implicitly as zero observations in the buy and sell sequences. With the introduction of the no-trade events, the duration between trades may provide information to market participants, because long durations between trades likely indicate the absence of news, since informed traders trade only when there is a signal, thus it is reasonable to assume that variations in the trading intensity are positively related to the bahaviour of informed traders.

Easley et al. (1997a) further distinguish between two different trade sizes, small and large trades.

Easley et al. (1997b) introduce the possibility that uninformed traders condition their trades on the observed order flow, thus inducing serial correlation in the observed trading process. They argue that uninformed traders will take into consideration the trading history in placing their orders.

3 The Basic Sequential Trade Model

In this section we describe the sequential trade model introduced by Easley et al. (1996), for those who are not familiar with the the PIN approach. This model can be viewed as a benchmark model for sequential trade models, since it captures all the essential feature of these types of models.

They set up a mixed discrete and continuous time sequential model of the trading process. In this framework trades arise because of the interaction of three types of economic agents: informed and uninformed traders and a risk-neutral, competitive market-maker. Trades depend on the arrival rates of informed and uninformed traders, which are governed by independent Poisson processes, and on the likelihood of the occurrence of three different types of information events (no news, good news and bad news) which are chosen by nature every day, before the first trade take place. In this model, the difference between bid and ask is due to asymmetric information of market participants about the occurrence of information events. Other components of the spread, such as those caused by maintaining large inventory imbalance, or by monopolistic exercising of market power by market maker, are left aside.

The trading period is a trading day. Trading days are indexed $i = 1, ..., I$ and, within a trading day, time is continuous, indexed by $t \in [0, T]$. Information events are independently distributed and occur with probability α. These events are good news with probability $1 - \delta$, or bad news with probability δ. After the end of trading on any day the full information value of the asset is realized.

Let $(V_i)_I^{i=1}$ be the random variable giving the value of the asset at the end of each trading day. During day i, if an information event occurs, the value of the asset conditional on good news is $\overline{V}_i$, on bad news is $\underline{V}_i$. The value of the asset if no news occurs is $V_i^* = \delta \underline{V}_i + (1 - \delta) \overline{V}_i$, (assuming $\underline{V}_i < V_i^* < \overline{V}_i$).

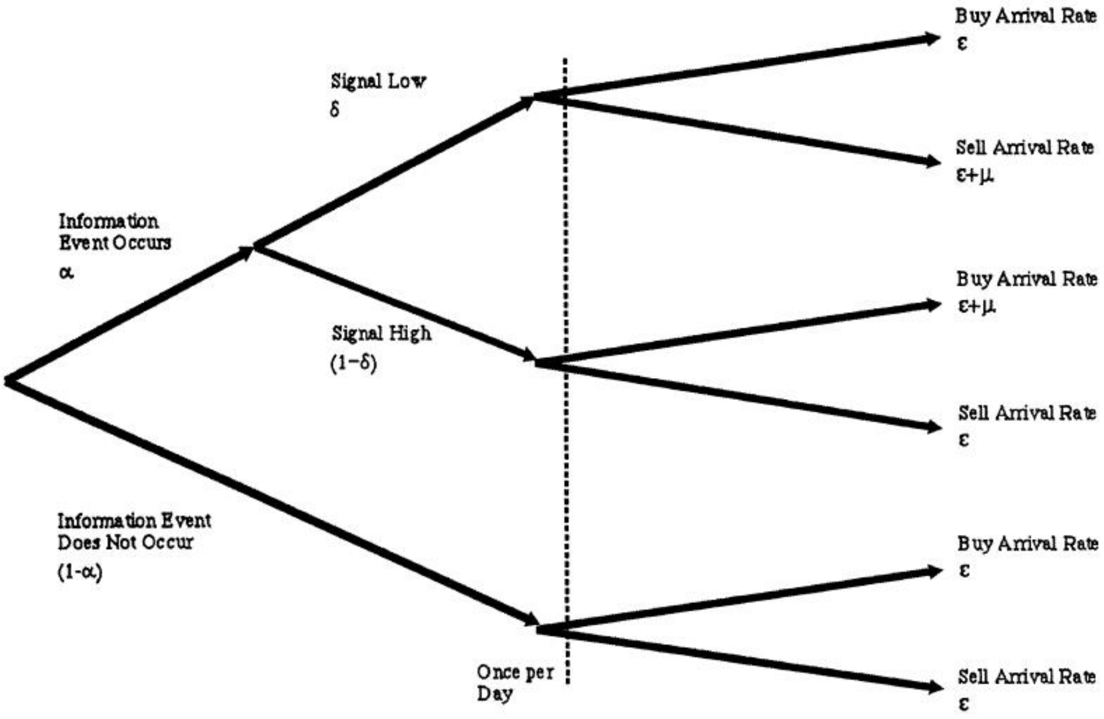
Trade arises from both informed traders (those who have seen any signal) and uninformed traders. On any day, arrivals of uninformed buyers and uninformed sellers are determined by independent Poisson processes. Uninformed buyers and sellers arrive at rate ϵ. The arrival rate of informed buyers and sellers is instead μ. Than, on good event days, the arrival rates are $\epsilon + \mu$ for buy orders and ϵ for sell orders. On bad event days, the arrival rates are ϵ for buys and $\epsilon + \mu$ for sells. Finally, on nonevent days, only uninformed traders arrive, and the arrival rate of both buys and sells is ϵ.

Each day nature selects one of the three branches of the tree. The market maker knows the probability attached to each branch, and he knows the order arrival process for each of the branches. He does not know, however, which of the three branches has been selected by nature. Since he cannot observe directly which type of information event has occurred, he used Bayes rule to update his beliefs about the nature of the information vent throughout the trading day.

Let $P(t) = (P_n(t), P_b(t), P_g(t))$ be the market maker's prior belief about the events no news (n), bad news (b), and good news (g) at time t. The unconditional priori beliefs at time 0 are equal to the probabilities with which nature chooses the information regime: $P_g(t) = \alpha \cdot (1 - \delta); P_b(t) = \alpha \cdot \delta; P_n(t) = 1 - \alpha$.

The tree in Figure 1 describes this trading process.

Figure 1. Tree diagram of the sequential trading process



Source: Easley et al. (1996)

After observing the first order arrival, posterior probabilities for all three events can be calculated recursively. Given that a sell order (S_t) arrives at time t, the probabilities are:

$$\begin{aligned} P_g(t|S_t) &= \frac{P_{g(t)\cdot\epsilon}}{\epsilon + P_{b(t)\cdot\mu}} \\ P_b(t|S_t) &= \frac{P_{b(t)\cdot(\epsilon+\mu)}}{\epsilon + P_{b(t)\cdot\mu}} \\ P_n(t|S_t) &= \frac{P_{n(t)\cdot\epsilon}}{\epsilon + P_{b(t)\cdot\mu}} \end{aligned} \tag{1}$$

The corresponding probabilities in case of a buy order (B_t) arrives at time t are:

$$\begin{aligned} P_g(t|B_t) &= \frac{P_{g(t)\cdot(\epsilon+\mu)}}{\epsilon + P_{g(t)\cdot\mu}} \\ P_b(t|B_t) &= \frac{P_{b(t)\cdot\epsilon}}{\epsilon + P_{g(t)\cdot\mu}} \\ P_n(t|B_t) &= \frac{P_{n(t)\cdot\epsilon}}{\epsilon + P_{g(t)\cdot\mu}} \end{aligned} \tag{2}$$

The bid-ask spread ($\Sigma(t)$) is given by the following relation:

$$\Sigma(t) = \frac{\mu \cdot P_g(t)}{\epsilon + \mu \cdot P_g(t)}(\overline{V}_i - E[V_i|t]) + \frac{\mu \cdot P_b(t)}{\epsilon + \mu \cdot P_b(t)}(E[V_i|t] - \underline{V}_i) \tag{3}$$

The first term of equation (3) is equal to the probability of observing information based buy order, multiplied by the expected loss of the market maker caused by such a transaction, while the second term reflects the expected loss caused by an information based sell order.

Thus, the initial belief of the market maker on the probability of information trading is:

$$PIN = \frac{\alpha\mu}{\alpha\mu + 2\epsilon} \tag{4}$$

and be interpreted as the unconditional share of informed trading.

Difficulties in estimating the parameter vector $\theta = (\alpha, \delta, \epsilon, \mu)$ arise because it is not possible to directly observe neither the occurrence of information events nor the associated arrival of informed and uninformed traders. However, knowing the daily arrives of sell (S_t)and busy (B_t) is it possible to infer these values using maximum likelihood and assuming that the trading process follows a Poisson process.

4 The VPIN Model

We use the new procedure developed by Easley et al. (2010b) to estimate the Probability of Informed Trading based on volume imbalance, i.e. the Volume-Synchronized Probability of Informed Trading (VPIN). This new approach overcomes the previous methodology issues related to the estimation of PIN for high-frequency trades, since it does not require the estimation of unobservable parameters $(\alpha, \delta, \epsilon, \mu)$ like explained in the previous paragraph. Moreover, the arrive of new information to the marketplace is updated in stochastic time.

They use the results of Easley et al. (2008). In this paper the authors consider the information content of total number of trades, because the numbers of buys and sells contain information about arrival rates of informed and uninformed traders. They calculate the expected arrival rate of informed trades, the traders than use this information to update their arrival rate estimates.

Defining $TT = S+B$ the total number of trades per day, the expected value of the total trades will be given by the sum of the Poisson arrival rates of informed and uninformed trades:
$E[TT] = \alpha(1-\delta)(\epsilon+\mu+\epsilon) + \alpha\delta(\mu+\epsilon+\epsilon) + (1-\alpha)(\epsilon+\epsilon) = \alpha\mu + 2\epsilon.$
The expected value of the trade imbalance ($K = S - B$) is given by: $E[K] = \alpha\mu(2\delta - 1)$. When the probability of bad news (δ) is not exactly one-half, the expected value of the trade imbalance provides information on the arrival of informed trades. The absolute value of the trade imbalance is: $E[|K|] \approx \alpha\mu$.
In Easley et al. (2010b) they use the same results, but considering the volumes. Indeed, the key idea behind the VPIN is that volume contains important pieces of information. When there is an important news there will be a large movement in the volume, so they divide

the data in volume ranges. They group trades into equal volume buckets of an exogenous size V. $\tau = 1, 2, ...n$ is the index of equal volume buckets. Within each volume bucket trades are classified as buys, V_τ^B, and sells,V_τ^S.

$V = V_\tau^B + V_\tau^S$ for each τ. Thus, the expected arrival rate of informed trade becomes: $E\left[V_\tau^S - V_\tau^B\right] = \alpha\mu(2\delta - 1)$, and the expected absolute value is: $E\left[|V_\tau^S - V_\tau^B|\right] = \alpha\mu$.

The expected arrival rate of total trades is:

$$\frac{1}{n}\sum_{\tau=1}^{n}(V_\tau^B + V_\tau^S) = V =$$

$$\underbrace{\alpha(1-\delta)(\epsilon+\mu+\epsilon)}_{\text{Volume from good news}} + \underbrace{\alpha\delta(\mu+\epsilon+\epsilon)}_{\text{Volume from bad news}} + \underbrace{(1-\alpha)(\epsilon+\epsilon)}_{\text{Volume from no news}} = \alpha\mu + 2\epsilon \tag{5}$$

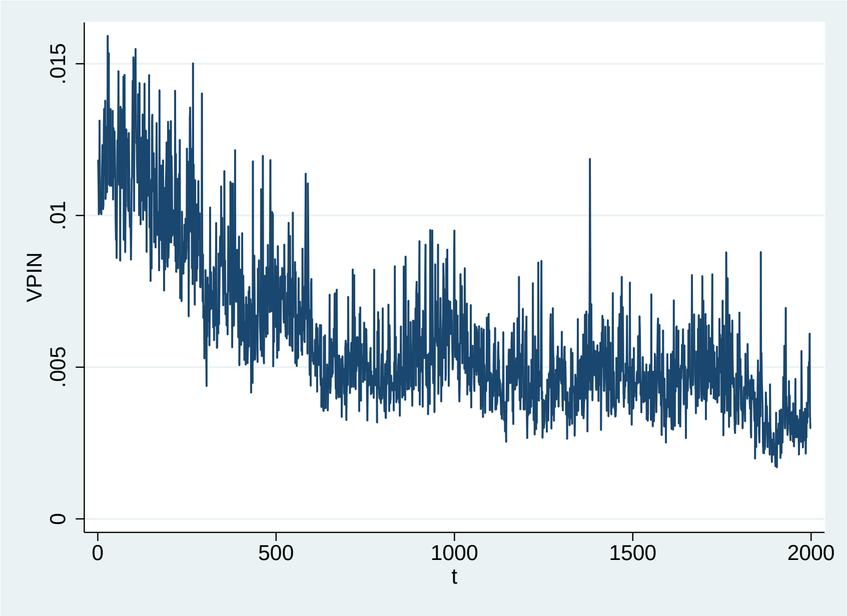
The Volume-Synchronized Probability of Informed Trading (VPIN) is computed in this way:

$$VPIN = \frac{\alpha\mu}{\alpha\mu + 2\epsilon} = \frac{\alpha\mu}{V} = \frac{\sum_{\tau=1}^{n}(V_\tau^S - V_\tau^B)}{(2\delta - 1)nV} \approx \frac{\sum_{\tau=1}^{n}|V_\tau^S - V_\tau^B|}{nV}{}^{1} \tag{6}$$

Microstructure models recognise that the arrive of orders contain important information, thus extracting information from order flow helps to determine subsequent prices. Easley and O'Hara (1992) already introduced the idea of the trading time instead of the clock time, arguing that the time elapsed between two trades gathered important information. Easley et al. (2010b)consider a new concept of trade time, as captured by volume. They suggest that more relevant an information is, more volume is attracted, so they sample the data on the base of volume.

[1] Look at the appendix in Easley et al. (2010b) for a detailed algorithm for the derivation of the VPIN.

Figure 2. VPIN metric



5 High Frequency Volatility Estimators and Jump detection

In this section we briefly described the volatility estimators used in our empirical analysis below, as well as the jump detection tests used to count the number of jumps we have in each day of our sample. We will use these estimators later on to investigate the link between VPIN and volatility.

We consider three alternative volatility estimators, estimators that employ the available high frequency data we have for SPY. The alternative volatility estimators we use take into account common problems arising in higher frequencies, i.e. biases induced by microstructure noise and price discontinuities.

The process we assume for the log price (X_t) process is the following:

$$X_t = \int_0^t \sigma dW_s + \int_0^t \kappa dN_s, \tag{7}$$

where κ is the random jump size and N_t a Poisson counting process with an adapted stochastic intensity parameter λ_t.

By computing the quadratic variation of X_t, i.e. the RV (realized variance), of equation

(7) it can be easily shown that in the presence of jumps the RV is a biased estimate of the true volatility ($[X,X]_T$).

$$RV^{(all)}_{X,T} = [X,X]_T + [J,J]_T = [X,X]_T + \sum_{i=1}^{N_T} \kappa^2_{\tau_i}, \tag{8}$$

where the quantity $\sum_{i=1}^{N_t} \kappa^2_{\tau_i}$ is the contribution of the jumps process to the $RV^{(all)}_{X,T}$. The superscript (all) means that we use all available data in our sample, something that we will not do in our analysis for the two out of the three volatility estimators (BPV - Bipower Variation, and $TBPV$ - Threshold Bipower Variation) because both estimators are designed to take into account only biases induced by the jumps and not by microstructure noise. We estimate both BPV and $TBPV$ using 5-min data. For the third estimator, MLE-F, we use min-by-min because it is a volatility parametric estimator that takes into account both microstructure noise and jump biases.

5.1 Non-parametric volatility estimators

To our knowledge, the first attempt to derive consistent estimates of the volatility σ of the Brownian part of the process X, in the presence of Poisson-type jumps, was that of *Power Variation*, introduced by Barndorff-Nielsen and Shephard (2004). The most widely used estimator that focuses on the continuous part of (8) is the well-known *Bipower Variation*, defined as

$$BPV_t = \mu^{-2} \sum_{i=2}^{N} |r_{i-1}||r_i|, \tag{9}$$

where r_i indicates the log-return, N is the total number of observations and $\mu \simeq 0.7979$.

The alternative volatility estimator we use in our analysis below is an extension of Multipower Variation which incorporates the concept of the threshold approach. The estimator called $TBPV$, *Threshold Bipower Variation*, proposed by Corsi et al. (2008) is given by

$$TBPV_{X,T} = \mu^{-2} \sum_{j=2}^{N} |r_{j-1}||r_j| \mathbf{1}[|r_{j-1}|^2 \leq \Theta_{j-1}] \mathbf{1}[|r_j|^2 \leq \Theta_j], \tag{10}$$

r_j is the log return, Θ_j the threshold function, $\mathbf{1}[\cdot]$ the indicator function, γ_k, $k = 1, \ldots, M$, are positive constants and $\mu = 0.7979$ as above. The idea is here is to disentangle the diffusion from jumps using the modulus of continuity of the Brownian motion. For this reason a suitable threshold is chosen, a threshold that vanishes slower than the modulus of continuity. The threshold we use is exactly the same as in Corsi et al. (2008).

The $TBPV$'s advantage is that it gives unbiased estimates of volatility when consecutive jumps appear in our price process. The simpler Multipower Variation is highly affected by the presence of consecutive jumps and the bias of the volatility estimator could be extremely large.

6 Jump detection tests and the MLE-F semi-parametric volatility estimator

6.1 Detecting jumps

Here, we review the Lee and Mykland (2008) and Lee and Hannig (2009) jump detection tests. Those jump detection tests will be used to calculate the exact number of jumps we have at each trading day of our sample and also to calculate the MLE-F volatility estimator we describe below.

Although both tests have been developed to be applied to high frequency data in the absence of microstructure noise, we also discuss that for practical purposes, depending on the variance of the microstructure noise σ_ε^2, both tests can be applied and jump detection is not affected by the presence of the microstructure noise.

6.1.1 Detecting Poisson-type jumps

Lee and Mykland (2008) propose a non-parametric test based on high frequency data to detect jumps that are generated by a non-homogeneous Poisson-type jump process. The test-statistic is based on the idea that if a jump occurred at time t_i, the return would be much larger than with usual innovations, while the *instantaneous volatility*, which in this case is an estimator not affected by jumps, would remain at the usual level. The statistic is

$$\mathcal{L}(i) := \frac{\log\left(\frac{S(t_i)}{S(t_{i-1})}\right)}{\widehat{\sigma}(t_i)}, \tag{11}$$

where the instantaneous volatility $\widehat{\sigma}(t_i)$ is given by

$$\widehat{\sigma}^2(t_i) := \frac{1}{K-2} \sum_{j=i-K+2}^{i-1} |\log\left(\frac{S(t_j)}{S(t_{j-1})}\right)||\log\left(\frac{S(t_{j-1})}{S(t_{j-2})}\right)|$$

where $S(t_i)$ denotes the stock price at time t_i.

Then, the ith observation is considered a jump if

$$\frac{|\mathcal{L}(i)| - C_n}{S_n} > 4.6001$$

where

$$C_n = \frac{(2\log n)^{1/2}}{c} - \frac{\log\pi + \log(\log n)}{2c(2\log n)^{1/2}}, \qquad S_n = \frac{1}{c(2\log n)^{1/2}},$$

$c = (2/\pi)^{1/2}$, n is the total number of observations and K is the time "window" used to calculate the instantaneous volatility.

6.1.2 Detecting infinite activity jumps from a Lévy jump process

Lee and Hannig (2009) extend the above non-parametric test to detect Lévy-type jumps: jumps that are difficult to locate due to their infinite activity and their small size which makes it difficult to differentiate them from price changes that are a result of a Gaussian shock.

Being able to detect all Lévy-type jumps in log prices allows us to separate the contribution to the noisy log-price $Y_t = X_t + \varepsilon_t$ that comes from a Gaussian shock or a jump. ε_t is the microstructure noise component which we assume to be normally distributed with mean zero and volatility σ_ε. This decomposition is crucial because we can reduce the empirical problem of estimating the daily volatility σ to one that we can solve by applying an MLE method to estimate volatility to the noisy series $\widetilde{Y}_t$, where $Y_t = X_t + \varepsilon_t$, .

6.1.3 *MLE-F* semi-parametric estimator

The *MLE-F* volatility estimator is designed to deal with microstructure noise and jumps in the log-prices proposed by Cartea and Karyampas (2009). It is a two-step approach. First, detect the returns in which a jump occurred and delete them (see Lee and Mykland (2008) and Lee and Hannig (2009) for jump detection). Second, once the de-jumped series is available, apply a maximum likelihood method to get fully efficient volatility estimates in the presence of microstructure noise, assuming, of course, that the distribution is correctly specified, i.e. Y_t is normally distributed. Note that deleting the jumps for the initial return series, makes the sample an irregularly spaced data set. This requires the maximum likelihood method used to estimate the volatility to incorporate this irregularity.

After having constructed the return series which do not contain possible discontinuities, we are left with $Y_t = X_t + \varepsilon_t$, where the log-price process X_t is a pure Brownian motion (σdW_t). The noise part can be "filtered" out using the same concept with Aït-Sahalia et al. (2005).

As we said above, the Δ variable is not always equal to $1/N$. For instance, when a return observation is deleted the next observation of our sample has a Δ value equal to $2/N$. Similarly, if two consecutive observations are jumps, according to the jump detection tests, then, after removing them, the time between the following observation and the previous one in the new $\tilde{r}_i$ series is $3/N$.

Hence, a non-constant Δ variable should be defined as $\Delta_i = \tau_i - \tau_{i-1}$. Once the Δ_i is computed for each observation we get fully efficient estimates for the volatility of the diffusion part and the variance of microstructure noise by maximizing the likelihood function:

$$l(\sigma^2, \sigma_\varepsilon^2) = -\ln\det(\Sigma)/2 - (N/2)\ln(2\pi) - \frac{1}{2}\tilde{r}'\Sigma^{-1}\tilde{r}, \tag{12}$$

where Σ is the covariance matrix of the returns:

$$\Sigma = \begin{pmatrix} \sigma^2\Delta_1 + 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & 0 & \cdots & 0 \\ -\sigma_\varepsilon^2 & \sigma^2\Delta_2 + 2\sigma_\varepsilon^2 & -\sigma_\varepsilon^2 & \ddots & \vdots \\ 0 & -\sigma_\varepsilon^2 & \sigma^2\Delta_3 + 2\sigma_\varepsilon^2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -\sigma_\varepsilon^2 \\ 0 & \cdots & 0 & -\sigma_\varepsilon^2 & \sigma^2\Delta_N + 2\sigma_\varepsilon^2 \end{pmatrix},$$

and $\tilde{r} = (\tilde{r}_1, \ldots, \tilde{r}_N)'$ is the returns vector that does not contain jumps. Although Aït-Sahalia et al. (2005) do not focus on the case where jumps are present in the price process, it provides an effective computer routine to estimate the above parameters when discussing irregularly spaced data, where the gaussianity assumption for the returns is still valid.

7 Data and Institutional Environment

The Exchanges Traded Funds We use data on the exchange traded fund tracking the S&P 500 index; the S&P 500 SPDR traded on Amex with ticker SPY and available on the TAQ database. This is not only considered from most financial historians the first ETF to come to the market in 1993[2], but it is also one of the most actively traded, regularly trading more than 100 million shares per day and sometimes reaching a peak of 400 million per day. The SPDR is structured as a unit investment trust, a mutual fund trust which does not require a board of directions. Moreover, it did not need an investment advisor, because the trustee build a mechanism into the trust so that its holding would match those of the S&P500 Index.

The choice of the data is due to the increasing importance of the ETFs in the investment horizon. During the last decade, the demand for ETFs has noticeably increased, as both institutional and retail investors progressively more rely on them to diversify the investment options in their portfolios and to hedge against broad movements in the stock market. With the increasing demand, also the number and the variety of investment objectives

[2] A variation of an exchange traded product was introduced in Canada in 1989 to track the largest stocks traded in Toronto, the Toronto Index Participation Fund, however most financial historians date back to 1993 the true beginning of the ETF industry, with the introduction of the SPDRs.

increased. We passed from a situation in which there was only one ETF in 1993 to the actual scenario where, according to the Investment Company Institute, by the end of 2009, the total number of index-based and actively managed ETFs had grown to 797, with total net assets of $777 billion.

The EFTs market increased rapidly since the launch of the SPDR in early 1990s. We assisted to an explosive pattern, from only two ETFs in 1995 to more than 1,000 in 2010. Furthermore, also the structure of the ETFs is changed through time. ETFs have evolved from an instrument mainly used by stock market professionals to one widely used by retail investors. The strategy of the ETFs is to offer a low-cost, tax-efficient, diversified way to invest. They pay fewer expenses to intermediaries (such as mutual funds managment companies) and reduce tax payments allowing investors keeping more of their money and, at the same time, they create diversified portfolios that meet the investor's needs.

Although the ETFs are similar to mutual funds, they differ for key operational and structural characteristics.

One of the main differences is that retail investors buy and sell ETF shares on a stock exchange through a broker-dealer, much like they would trade any other type of stock.

Another major discrepancy between mutual funds and ETFs is pricing. Mutual funds are forward priced, in the sense that while investors can place orders to buy or sell shares through the day, all orders placed during the day will receive the same price once per day at the end of that day's trading (usually at 4:00pm EST, the U.S. stock exchanges closing time).

There are many characteristics that can explain the popularity of the ETFs, with respect to other instruments, like mutual funds (Maeda, 2008), the most important are:

Diversification– an ETF is composed of securities selected from a particular index so it is easy to diversify a portfolio, decreasing the risk with respect to placing the capital in the shares of a single company.

Transparency– an ETF is required to make available at all times information about the underlying stocks that it holds, so the investors can easily know, at any time, which stocks are included in the fund and what fees and expenses are incurred by the fund.

Instead mutual funds are vague about their holdings because fund managers do not want to publicly disclose their investment strategies[3]

Lower fees and commissions.– There are lower fees and commissions with respect to a mutual fund. Since the shares of an ETF are not sold directly by the company, but they are through a broker on the open market. In order to buy or sell shares in shares in an ETF, an investor pay a single, low transaction fee to the brokerage. Instead, many mutual funds, in addition to commissions, have also load and exit fees that investors pay to enter or exit the fund which must be subtracted to returns from investment. Moreover, because an ETF is linked to an index and it is not actively managed, there is only a license fee to pay to the the provider of the index, such as S&P.

Tax efficiency– ETFs are tax-efficient because they are structured to minimise capital gains tax. If a mutual fund sells some of its assets for profit, the individual investors in that fund are required to pay capital gains tax even if the fund as a whole loses value during the year. This is something that does not happen with an ETF because securities are efficiently owned by a third party.

Ease of purchase– ETF shares can be bought and sold anywhere, and shares of ETFs from different providers can be managed in a single brokerage account, instead, investors who wish to hold more than one index mutual fund has to maintain several separate accounts. Thus, ETFs offer the possibility to invest in an entire market sector with one trade and to access to market sectors and indexes that mutual funds do not currently offer.

Investor strategies– Shares of a ETFs are sold on the stock exchanges throughout the day and can be employed in all the investment strategies that are used with stocks, such as options or hedging.

The winning strategy of the ETF is to offer a low-cost, tax-efficient, diversified way to invest. They pay fewer expenses to intermediaries (such as mutual funds management companies) and reduce tax payments allowing investors keeping more of their money and, at the same time, they create diversified portfolios that meet the investor's needs.

[3] Sometimes can occur the phenomenon of the ''style drift'', i.e. over time, fund managers can change the character of their fund, without investors be aware of this. In particular, the SPDRs which are organised as unit investment trust are strictly required to maintain the holdings that reflect the composition of the index they are tacking.

7.1 Trade Direction Algorithms

The TAQ database provides a complete list of quotes, trades and volumes at each point in time for each day. Our sample period goes from January 2000 to December 2009, for a total of 1998 trading days. An important piece of information for carrying out our analysis is the trade direction, i.e. whether a trade was buyer or seller initiated. However, the TAQ database records only transactions but not the party who initiated the trade. In order to classify a trade as buyer or seller initiated we need to use one of the classification techniques developed in the literature. The widely used technique to ascertain the trade direction is the so called Lee-Ready algorithm after Lee and Ready (1991), who originally proposed the method. Other commonly accepted trade direction algorithms include the tick test and the quotes test[4].

The tick test classifies the current trade as buyer (seller) initiated, if the price of the previous trade is lower (higher) than the price of the current one. If the price of two consecutive trades is the same, then the last recorded price change is used to establish the direction. This method is very simple to implement since the only indispensable information is the sequences of prices. However, it is unable to classify transactions at the beginning of the day and trades priced at the midquote between the bid and the ask.

The quote test classifies a trade as buyer (seller) initiated, if the current trade is closer to the last recorded ask (bid) quote. If the current price is at the midquote, the last recorded price change is used to determine the direction. The Lee-Ready algorithm is usually recognised as the most accurate, nevertheless the degree of accuracy ranges from 72% to 93%, depending on the study (Asquith et al., 2010). It combines trades and quotes to establish trade direction and can be divided in the following steps:

1. If the current quote has not been changed within the last 5 seconds before the transaction, we determine the trade direction comparing the current quote to the trade price. When the price is equal to the ask (bid), the trade is classified as a buy (sell).

2. If the current quote is less than 5 second old, the trade price is compared to the previous quote.

[4] For further information on the trade classification algorithms, see, for example, Bessembinder (2003)

3. If the trade prices, as determined in step one or two, are outside the spread, the trade is classified using the distance of the price form the closest quote. When the price is greater than the ask (smaller than the bid), the trade is classified as a buy (sell).

4. If the price is at the midquote, we use the tick test to classify the trade.

5. If the price is inside the spread, but not at the midquote, the trade is classified considering its distance to the closest quote. When the price is closer to the ask (bid), the trade is classified as a buy (sell).

6. If neither of the previous conditions applies, the trade direction is indeterminable.

Furthermore, Lee and Ready (1991) highlight a potential problem in matching the trades with the quotes. The prevailing quote at the time of a trade may not be that trade's associated quote. They address the problem excluding quotes that occur less than 5 seconds before the current trade.

In classifying the trades in our dataset we take into consideration this rule of the 5 seconds delay, moreover, for completeness, we also consider the 1 second delay rule (Henker and Wang (2006)) and the zero second delay rule; trades are compared to quotes in effect at the trade report time, without any adjustment to time stamps to allow for possible reporting delays (Ellis et al. (2000), Bessembinder (2003), Peterson and Sirri (2003)), to see if there are significant differences in the classification. We are aware that neither of these methods is exempt from errors, however, we need to rely upon rules of inferences and empirical studies conclude that they are reasonably accurate. Moreover, Asquith et al. (2010) emphasise that all the rules of inference apply to single-sided market or executable limit orders, but some datasets (like TAQ) do not identify order type, potentially increasing the error rate. In addition, tests of the trading algorithms occurred before the decimalisation in 2001, which has led to the narrower bid-ask spreads, making trade classification more difficult.

8 VPIN and Volatility

Easley et al. (2010b) do not explicitly propose the VPIN to forecast volatility, since VPIN uses only volume and not price information. However, we can expect a close relation between the VPIN and the volatility. Indeed if it is true that large price movements are associated with large volumes, sampling by volume is a proxy of sampling by volatility. This transformation allows also a partial recovery of normality. The historical distribution of VPIN for our data mimic the one in Easley et al. (2010a) and can be approximated by a long-normal distribution. (see Figure 3).

However, the measure of volatility used by Easley et al. (2010b), i.e. the absolute value of the returns, is not the most efficient estimator of volatility. We will use volatility estimators described in section 5 and 6.

Easley et al. (2010b) find a positive correlation between the VPIN and their measure of volatility. During period of high VPIN we would observe an increase in volatility as a result of liquidity providers withdrawing from the market. We find a very large positive and statistically significant correlation (0.3292) between the VPIN and the future volatility.

We also investigate whether the VPIN Granger-causes our volatility measure or whether volatility Granger-causes VPIN. We find results comparable with those that Easley et al. (2010b) find for the futures contracts, i.e. that the casual link goes from VPIN to volatility.(see Table 2 in the Appendix).

We further investigate the relation between VPIN and volatility using a Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) introduced by Corsi (2009). He shows that this model reproduces all the principal features of financial returns. We use the model in the version presented by Corsi and Renò (2010). They add to the lagged volatitlites (at daily, weekly and monthly frequencies), the jumps as another possible source of future volatility. They show that volatility increases after a jump, but that this shock is absorbed quickly. We modify this model introducing the VPIN, to draw some conclusions about its relation with the volatility.

Table 1. HAR Estimation with Number of Jumps and VPIN

$RV^{(d)}_{t+1d} = c + \alpha_0 RV^{(d)}_t + \alpha_1 RV^{(w)}_t + \alpha_2 RV^{(m)}_t + \beta_0 NJ^{(d)}_{t+1d} + \gamma_0 VPIN^{(d)}_t + \gamma_1 VPIN^{(d)}_{t+1}$

	Model 1 (TBPV)	Model 2(MLE-F)	Model 3(TBPV)	Model 4(MLE-F)	Model 5(TBPV)	Model 6 (MLE-F)
c	0.626	0.534	0.888	0.795	−0.663	−0.019
	(3.425)	(2.293)	(3.086)	(4.074)	(−3.111)	(−0.117)
α_0	0.418	0.544	0.426	0.556	0.321	0.488
	(4.415)	(16.644)	(4.419)	(17.628)	(3.241)	(14.223)
α_1	0.421	0.319	0.412	0.307	0.368	0.293
	(3.513)	(5.278)	(3.547)	(5.409)	(4.800)	(5.658)
α_2	0.123	0.102	0.125	0.103	0.255	0.174
	(3.341)	(2.827)	(3.606)	(2.982)	(4.572)	(5.100)
β_0					1.065	0.560
					(5.559)	(7.547)
γ_0			−417.11	−415.224	−247.22	−326.247
			(−5.595)	(−6.52)	(−4.905)	(−6.486)
γ_1			360.348	362.357	277.74	316.056
			(5.836)	(6.611)	(5.673)	(6.586)
AdjR^2	0.7999	0.8358	0.8031	0.8402	0.8800	0.8681

t statistics in parentheses

$$RV^{(d)}_{t+1d} = c + \alpha_0 RV^{(d)}_t + \alpha_1 RV^{(w)}_t + \alpha_2 RV^{(m)}_t + \beta_0 NJ^{(d)}_{t+1d} + \gamma_0 VPIN^{(d)}_t + \gamma_1 VPIN^{(d)}_{t+1} \quad (13)$$

where $RV^{(d)}_t$, $RV^{(w)}_t$, and $RV^{(m)}_t$ are respectively the daily, weekly and monthly observed realized volatilities; NJ is the number of jumps and VPIN is the Volume-Synchronized Probability of Informed Trading.

We than check for other models, imposing to the equation (13) the following restrictions: $\beta_0 = 0$ and $\beta_0; \gamma_{0,1} = 0$. The results are shown in Table 2.

9 Conclusions

We use the new procedure developed by Easley et al. (2010b) to estimate the Probability of Informed Trading. Instead of applying maximum likelihood procedure to estimate unknown parameters we use the volumes as a proxy for the arrive of new information. The idea is that more relevant is a piece of information, more volume it will attract. Moreover, the VPIN can be seen also as a proxy of volatility since large price movements are associated with large volumes, thus sampling by volume is a proxy to sampling by volatility.

We use data on the exchange traded fund tracking the S& P 500 index, the SPY, for a period of ten years. The choice of the data is due to the increasing importance of ETFs in the investment horizon. Our results are comparable to those in Easley et al. (2010b). We find that the historical distribution of Probability of informed trading follows also in our case a log-normal distribution and also the trend of our VPIN is comparable to the one for the futures. We also investigate whether VPIN metric Granger-cause volatility of whether the reverse is true. We find support that the casual link goes from VPIN metric to volatility, i.e. VPIN improves the forecast of volatility.

We further investigate the relation between VPIN and volatility using a Heterogenous Autoregressive model of Realized Volatility with jumps like in Corsi and Renò (2010) and adding the VPIN. We compare different models with and without jumps and VPIN and we find that the model that includes both jumps and VPIN is better than the others.

As a further investigation we plan to analyse the possible links between VPIN and liquidity measures.

References

Aït-Sahalia, Y., Mykland, P., and Zhang, L. (2005). How often to sample a continuous-time process in the presence of marketmicrostructure noise. *Review of Financial Studies*, 18(2):351–416.

Asquith, P., Oman, R., and Safaya, C. (2010). Short sales and trade classification algorithms. *Journal of Financial Markets*, 13(1):157–173.

Barndorff-Nielsen, O. and Shephard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1):1–37.

Bessembinder, H. (2003). Issues in assessing trade execution costs. *Journal of Financial Markets*, 6(3):233–257.

Cartea, Á. and Karyampas, D. (2009). The relationship between the volatility of returns and the number of jumps in financial markets. *Open Access publications from Universidad Carlos III de Madrid*.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*.

Corsi, F., Pirino, D., and Renò, R. (2008). Volatility forecasting: the jumps do matter. *working paper*.

Corsi, F. and Renò, R. (2010). Discrete&time volatility forecasting with persistent leverage effect andthe link with continuous&time volatility modeling. Technical report, Working paper.

Easley, D., de Prado, M., and O'Hara, M. (2010a). The microstructure of the flash crash: Flow toxicity, liquidity crashesand the probability of informed trading.

Easley, D., de Prado, M. L., and M.O'Hara (2010b). Measuring flow toxicity in a high frequency world.

Easley, D., Engle, R., O'Hara, M., and Wu, L. (2008). Time-varying arrival rates of informed and uninformed trades. *Journal of Financial Econometrics*, 6(2):171.

Easley, D., Kiefer, N., and O'Hara, M. (1997a). The information content of the trading process. *Journal of Empirical Finance*, 4:159–186.

Easley, D., Kiefer, N., and O'Hara, M. (1997b). One day in the life of a very common stock. *Review of Financial Studies*, 10(3):805–835.

Easley, D., Kiefer, N., O'Hara, M., and Paperman, J. (1996). Liquidity, Information, and Infrequently Traded Stocks. *Journal of Finance*, 51:1405–1436.

Easley, D. and O'Hara, M. (1987). Price, trade size, and information in securities markets. *Journal of Financial Economics*, 19(1):69–90.

Easley, D. and O'Hara, M. (1992). Time and the process of security price adjustment. *Journal of Finance*, 47(2):577–605.

Ellis, K., Michaely, R., and O'Hara, M. (2000). The accuracy of trade classification rules: Evidence from Nasdaq. *Journal of Financial and Quantitative Analysis*, 35(04):529–551.

Glosten, L. and Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneouslyinformed traders. *Journal of Financial Econometrics*, (14):71–100.

Henker, T. and Wang, J. (2006). On the importance of timing specifications in market microstructure research. *Journal of Financial Markets*, 9(2):162–179.

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

Lee, C. and Ready, M. (1991). Inferring trade direction from intraday data. *Journal of Finance*, 46(2):733–746.

Lee, S. and Hannig, J. (2009). Detecting jumps from Lévy jump diffusion processes. *Forthcoming in Journal of Financial Economics.*

Lee, S. and Mykland, P. (2008). Jumps in financial markets: A new nonparametric test and jump dynamics. *Review of Financial studies*, 21(6):2535.

Maeda, M. (2008). *The Complete Guide to Investing in Exchange Traded Funds: How to Earn HighRates of Return-Safely.* Atlantic Publishing Company (FL).

O'Hara, M. (1998). *Market microstructure theory.* Wiley.

Peterson, M. and Sirri, E. (2003). Evaluation of the biases in execution cost estimation using trade and quotedata. *Journal of Financial Markets*, 6(3):259–280.

Appendix

Figure 3. VPIN metric and a log-normal distribution

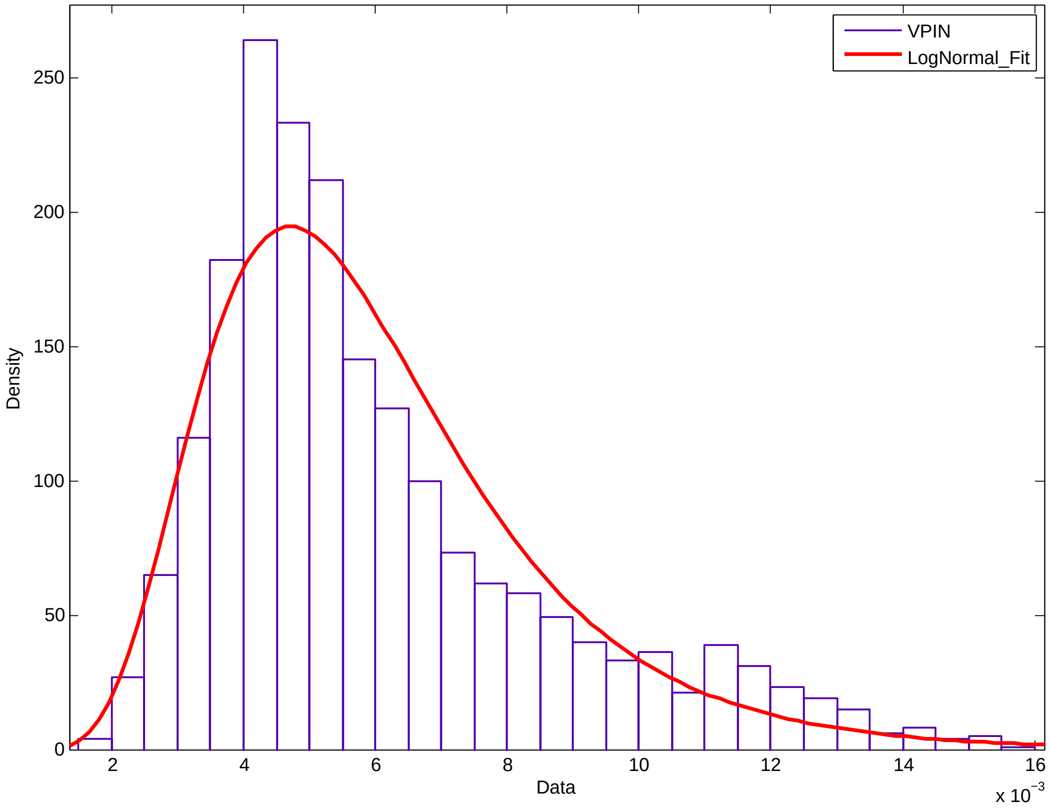


Table 2. Granger-causality

	$\mathrm{Y}_t = TBPV$	$\mathrm{Y}_t = MLE - F$	$\mathrm{Y}_t = VPIN$	
Intercept	1.52558	1.6524	0.0011	
	(5.79)	(7.58)	(13.24)	
Y_{t-1}	0.8941	0.8769	0.8082	
	(42.32)	(45.89)	(50.68)	
AdjR^2	0.7996	0.7693	0.6543	
	$\mathrm{Y}_t = TBPV$	$\mathrm{Y}_t = MLE - F$	$\mathrm{Y}_t = VPIN$	$\mathrm{Y}_t = VPIN$
Intercept	0.5341	0.7909	0.0008	0.008
	(2.07)	(2.96)	(8.50)	(9.33)
TBPV_{t-1}	0.8598		0.001	
	(37.24)			(8.40)
$\mathrm{MLE\text{-}F}_{t-1}$	0.8718	0.8520	0.0004	
	(43.87)	(43.50)	(7.27)	
VPIN_{t-1}	247.0576	198.6759	0.7734	0.7541
	(5.90)	(0.94)	(46.94)	(44.32)
AdjR^2	0.8051	0.7744	0.6643	0.6681

t-statistics in parentheses