Evaluating market's contribution to price discovery in High-Frequency for Co-listed assets*

Christian K. Nguenang†
Toulouse School of Economics

Mai 2016

Abstract

Assets are traded on many places that are remotely situated. An outstanding question is how those places individually contribute to price discovery. We provide way to study this problem in the context of high frequency data. We propose a measure evaluating the permanent impact of a shock on a market's innovation. It has advantages on the literature in that: it is in continuous-time and deals with non-informative microstructure noise; it provides a unique meaningful measure of information processing. We provide an application to some SP500 stocks co-listed on NYSE, NASDAQ and some regional exchanges. The results suggest that the price discovery mostly happens on NYSE for big and medium size trade, on NASDAQ for small size trade, the contribution of a market is correlated with its market share for small and medium size transactions and it is not large quantities trades that convey information.

Keywords: *continuous-time cointegration, Generalised Impulse Response, preaveraging, Price discovery, Modulated Realized Covariance*

JEL: *C32, C58, G14*

*We thank Nour Meddahi for guidance and supervision. We are also grateful to Fanny Declerk and Sophie Moinas for helpful discussions and comments. Views expressed are those of the author

†Tel: +33 9 07 30 69 04 70 88, Toulouse School of Economics (TSE). 21 Allée de Brienne 31000 Toulouse - FRANCE. E-mail: Christian.Nguenang-kapnang@ut-capitole.fr;

Contents

1 Introduction

The perpetual improvements in financial markets and development of High frequency trading (HFT) is accompanied by a growing academic literature on the resulting consequences on market quality. There is an ongoing integration of financial market at the national, and at the international level. Assets are traded on multiple market places and traders can similarly send orders for the same security in many platforms, which can be geographically far one from another. Among market's characteristics, the price discovery mechanism should deserve more attention. There is clearly an interest to understand and to evaluate how markets compute new information into prices. Understanding this process should be valuable for an international investor, when designing his trading strategy, or when choosing how to split the orders in different market places. The regulator also in its quest to the best market organization could find important to know which market contributes the most to the price movement of an asset[1]

For multiple market study of price discovery, some measures were developed in the literature triggered by Hasbrouck (1995). He presented a measure of price discovery called Information Share (IS) and provides comparison of New York Stock Exchange and the Regional exchanges in the quotes formation of thirty Dow stocks. The main competing measure is the common factor weight of Gonzalo and Granger (1995) permanent-transitory (PT) decomposition (see Harris et al., 2002b). Those methods are intensively discussed by De Jong (2002), Lehmann (2002), Hasbrouck (2002),Baillie et al. (2002),Yan and Zivot (2010). The main lesson is that the IS is more concerned with the variability in the process with an economic sensitive identification of its efficient price. The IS measure suggests to evaluate the market contribution to price discovery by the relative part of this market in the variance of the innovation in the efficient price. Meanwhile IS has some drawbacks; it is not identified and is only able to produce upper and lower bound, in addition it assumes a one-to-one cointegrating relation, so it is well suited for strongly related asset like the same stock traded on two market places. But when it comes to study the spot and futures market on a given stock, the assumption that the cointegrating coefficient is one is questionable (see Lien and Shrestha (2014)). It is recommended to sample at high frequency to reduce the correlation and then tighten the IS bounds, but this practice ignore that at high frequency non informative part of the noise dominate the variances estimation[2]

Many authors tried to solve this identification issue by doing some transformations of the

[1] Eun and Sabherwal (2003a) report that the Canadian authority was really worried about US-markets becoming the place where the Canadian's stock prices were computed

[2] This is related to the signature plot of Andersen et al. (2000)

innovation variance matrix. But the limit of those techniques is that they completely lost an economic meaning behind the mathematical operations. For example Lien and Shrestha (2014) use an orthogonalization of the correlation matrix to propose the Generalized Information Share (GIS), this measure has the advantage of being independent of the variables ordering and is applicable to CDS and Bond as in their application. Meanwhile the orthogonalization procedure of the correlation matrix lacks some economic intuition. This is the same problem of method based on identification through heteroskedasticity as in Grammig and Peter (2013).

On the pure applied side, some papers use the model-free price discovery measures provided by the previous papers to study the determinant of market performance. They are interested in which variables explain one market's contribution to price discovery mechanism: Blume and Goldstein (1997); Chakravarty et al. (2004); Huang (2002), Barclay et al. (2003).

Even if they differ in the way they define price discovery mechanism, all those studies identify their price discovery measure by relying on an error correction model of the non stationary price process. In addition their construction is based on a model that considers only microstructure noise related to information sources: Information asymmetry, market under/over reaction (Menkveld et al., 2007). It is not concerned with non-informative noise due for example to measurement errors, bid-ask spread...etc.

Relying on Hasbrouck (1995) identification of the unobserved efficient price we measure market contribution by accessing how this permanent unobserved price reacts to a shock in one market. This is related to the generalized impulse response of Pesaran and Shin (1998) in the VAR literature, but here instead of looking the response function of each market, we look at the response function of the permanent common component to the markets. We show in a discrete time framework that it is a sensible way of defining price discovery with the advantage of being the first price discovery measure based on Hasbrouck (1995) efficient price to provide a testing framework. We then propose the **High-Frequency Information Share (HFIS)** that is best defined as ***the variance of the change in the fundamental price over a period of time, resulting from a shock to the innovation in one market*** . Our framework has the following advantages over the literature:

- It provides a unique value while maintaining an economic sensitive definition
- It uses a continuous time-setup to deal with high frequency data.
- It accommodates a stochastic volatility, important for example to capture clustering effects.

- It explicitly deals with non-informative part of microstructure noise, which is ignored in the literature.

The remainder of the paper is organized as follow: The second section presents a review of the principals existing measure of price discovery and finish by a new measure in discrete time; The third section presents the high frequency measure of price discovery in continuous time framework. In fourth section the estimation strategy is discussed and Monte-carlo simulation are discussed. In the fifth section an application is done on some SP500 stocks that are traded on NYSE, NASDAQ, and other regional exchanges.

2 Measuring price discovery in discrete time

2.1 Setup

There exist d strongly related securities that are traded at the respective prices $p_{1t}\dots$ and p_{dt}. For example if it is one asset listed on two markets, p_{1t} is the price of the asset on the first market and p_{2t} the price on the second. We denote the vector of prices by $P_t = (p_{1t}, p_{2t}\dots p_{dt})'$. In this situation it is classical that the price P_t is assumed to be cointegrated and the gap between every two prices is stationary such that there exist only one common trend for all the prices. Using Johansen (1991) results, P_t can be shown to admit the following Vector error correction model (VECM) representation:

$$\Delta P_t = -\alpha\beta' P_{t-1} + \Gamma_1 \Delta P_{t-1} + \ldots + \Gamma_K \Delta P_{t-K} + \varepsilon_t \tag{1}$$

where the cointegrating matrix is normalized here to be

$$\beta' = \begin{bmatrix} 1 & -\beta_1 & \dots & 0 & 0 \\ 0 & 1 & -\beta_2 & & 0 \\ & & \dots & & \\ 0 & 0 & & 1 & -\beta_{d-1} \end{bmatrix} : (d-1)\times d \tag{2}$$

The infinite moving average representation of the price vector difference and the Granger representation theorem gives the following relationships where $\Psi(L)$ is a lag polynomial and ε_t independent white noise with $var(\varepsilon_t) = \Omega$:

$$\Delta P_t = \Psi(L)\varepsilon_t = \left(\Psi(1) + \Psi^*(L)(1-L)\right)\varepsilon_t \tag{3}$$

$$P_t = P_0 + \Psi(1) \sum_{s=1}^{t} \varepsilon_s + \Psi^*(L)\varepsilon_t \tag{4}$$

The matrix of the long run impact is given by

$$\Psi(1) = \beta_\perp \left(\alpha_\perp' \left(I - \sum_{i=1}^{p} \Gamma_i \right) \beta_\perp \right)^{-1} \alpha_\perp' \tag{5}$$

From now on, to simplify the presentation and keep the focus on intuitions and arguments, we restrict to $d = 2$ markets. Meanwhile the proof are done for $d > 2$.

2.2 A review of price discovery metrics

2.2.1 The Information Share measure

For an asset that is traded on two or more venues, Hasbrouck (1995) is looking for a measure that will determine on which market the price discovery does happen. He proposed to use the contribution of each market in the variance of the innovation of the fundamental value (or the "efficient price" which is common to all the markets). His method relies on the assumption that the cointegrating equation is $\beta = (1 \ -1)$.

So the matrix $\Psi(1)$ can be expressed with a the unique ψ as

$$\Psi(1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \psi = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} \psi_{11} & \psi_{12} \end{pmatrix}$$

Replacing in equation 4 yields

$$P_t = P_0 + \begin{pmatrix} 1 \\ 1 \end{pmatrix} \psi \sum_{s=1}^{t} \varepsilon_s + \Psi^*(L)\varepsilon_t \tag{6}$$

The random walk component of the price is $\psi \sum_{s=1}^{t} \varepsilon_s$ it is a scalar random variable and it is common to market 1 and market 2. It is identified as the implicit fundamental price of the asset. The new information entering the fundamental price is $\psi\varepsilon_t$ and its variance $(\psi\Omega\psi')$ is the total information share. he defines the market contribution in the following way:

If Ω were to be diagonal, then the total information share will be

$$\psi\Omega\psi' = \psi_{11}^2\Omega_{11} + \psi_{22}^2\Omega_{22}$$

and the information share of the market j will be

$$IS_j = \frac{\psi_{jj}^2 \Omega_{jj}}{\psi \Omega \psi'}$$

As Ω is not diagonal in general, its Cholesky decomposition root is computed as $\Omega = FF'$ with F lower triangular. and we have

$$IS_j = \left([\psi F]_j\right)^2 / \psi \Omega \psi' \tag{7}$$

$[\psi F]_j$ is the jth element of the matrix ψF.

An identification problem arises because this decomposition implies an impact of the ranking of the variables on the result. Then by switching variables position in the price vector, he provides lower and upper bounds on the Information Share of each market.

2.2.2 Invariant information Share measures

The IS identification problem can be summarize in the structural shock identification problem in the SVAR literature. In fact having the reduced form shock ε_t the goal is to look for structural shock η_t and B such that $\eta_t = B^{-1}\epsilon_t$. The problem is summarize by solving for the matrix B in equation $\Omega = BB'$. Unfortunately and infinity of solutions exist and the Hasbrouck (1995) choice is to consider the lower triangular matrix F obtained from the Cholesky root of Ω. But as seen previously the IS doesn't give the same value for a market if it is placed in first and in second position in the price vector.

A solution to this is presented in Lien and Shrestha (2014) where instead of focusing on the covariance matrix Ω, they consider the eigenvalues decomposition of its correlation matrix Φ. Let G the matrix of eigenvectors, Λ is the diagonal matrix of eigenvalues and $V = diag\left(\Omega_{11}, \Omega_{22}\right)$ the diagonal matrix of standard deviations, and

$$F^* = \left[G\Lambda^{-1/2}G^T V^{-1}\right]^{-1} \tag{8}$$

It happens that $\Omega = F^* \left(F^*\right)^T$. They thus define their Generalized Information share for market j using the matrix $B = F^*$

$$GIS_j = \left([\psi F^*]_j\right)^2 / \psi \Omega^2 \psi$$

Where ψ is a line of the matrix $\Psi\left(1\right)$[3].

[3]Here they dont assume the Cointegrating value to be -1 like for the Modified Information Share (MIS)

This method has the advantage of being independent of the variables ordering, but it strongly lack an economic relevance behind the decomposition of the correlation matrix.

Others attempts to compute a unique Information Share may be to use non-gaussianity or heteroskedasticity to identify structural shocks as it is done in the Macroeconomics literature. Those procedures have the advantage of allowing identification of the two structural shocks. Meanwhile it is not possible to say which shock comes from which market and the parameters are identified only up to a permutation matrix. In addition to the fact that they are purely statistical identification schemes with no economics motivation, this is a severe problem for the purpose of assigning market contribution to price discovery. To overcome this problem, Grammig and Peter (2013) after considering heteroskedasticity on two regimes of structural innovations to identify uniquely the matrix B, they assign shocks in a way that the coefficient of a shock on its market should be bigger than its coefficient on the other market.

Another non less important disadvantage of method based on information share is that they lack asymptotic theory and testing. The current practice is to use some bootstrap procedures to provide standard errors on Information Share.

2.2.3 α-based measures

The main competitor to IS measure in the literature is the Gonzalo and Granger (1995) common factor weight in the Permanent-Transitory (PT) decomposition. This consist of decomposing a difference stationary time series as the sum of a permanent I(1) component Q_t and a transitory stationary component T_t. The identification of the two components of $P_t = Q_t + T_t$ relies on two assumptions:

- Q_t and T_t form a PT decomposition,
- T_t is a linear combination of the observed variables,

In the contest of one asset and many markets the permanent component is driven by an $I(1)$ factor (f_t) common to both markets such that the observed price vector can be written as

$$P_t = \begin{bmatrix} 1 \\ 1 \end{bmatrix} f_t + T_t$$

And it is shown that given the ECM equation 1, the weight in the $I(1)$ component are proportional to $\alpha_\perp$ such that:

in Lien and Shrestha (2009) . But GIS and MIS are analytically exactly equal, the only difference remains in the estimation of the GIS where there is no constraints on the coefficients β_1.

$$f_t = c \times \alpha_\perp P_t = c \begin{pmatrix} \alpha_{1\perp} & \alpha_{2\perp} \end{pmatrix} P_t \quad \text{,with } c \text{ constant}$$

The relative contribution to price discovery of market 1 and market 2 is thus computed by taking the weight of each market in the permanent component (Harris et al., 2002a) as

$$PT_1 = \frac{\alpha_{1\perp}}{\alpha_{1\perp} + \alpha_{2\perp}} \text{ , } PT_2 = \frac{\alpha_{1\perp}}{\alpha_{1\perp} + \alpha_{2\perp}}$$

A difference of the PT measure with the IS measure is that f_t is a linear combination of only the current prices. Thus the permanent component of the Gonzalo and Granger (1995) decomposition is generally not a random walk. This is a serious drawback as this permanent component could not represent an efficient price. Baillie et al. (2002) shows that both can be computed easily after the estimation of the ECM and they present the relationship linking CS to IS.

Instead of focusing on the innovation variation, the permanent component Share relies on the error correction weighting matrix $\alpha_\perp$. In this respect Eun and Sabherwal (2003b) also think of price discovery as the adjustment to the equilibrium but access price discovery of a market directly by its coefficient in α summarizing its speed of adjustment toward the long run equilibrium. Building the measures with only a coefficient of the VECM allows those methods to have testable implications and thus statistical significance checking of the contribution to price discovery.

2.3 New approach: The Impulse response Information share measure (IRIS)

Price discovery is defined as the process of compounding information into prices. For the cross listed assets the question of measuring price discovery is related to the desire to know on which market information enter the prices. As information is supposed to affect permanently the prices, an appealing intuition is to say that: if information comes trough market 1 and not through market 2 then the permanent price should react to innovation in market 1 and not to innovation in market 2. Price discovery can then be well evaluated by the response of the efficient price to each market's innovation. Even if this formulation stipulate a kind of impulse response function, it is different from the analysis of Yan and Zivot (2010) where after assuming two structural shocks (informational shock and a noisy shock), they are interested to the reaction of IS and CS to the informational shock. Here after identifying the same permanent price as Hasbrouck (1995) we are interested in the response of this efficient price

to an innovation shock in each market. This framework has the advantage of giving a unique value without loosing economic relevance and provides a rationale to IS Upper bound. In addition it is the first in this literature to provide testable results for price discovery measures based on Hasbrouck (1995) efficient price, .

2.3.1 The Generalized impulse response

In the case of linear VAR and Cointegrated systems, Pesaran and Shin (1998) analyzed the generalized impulse response function (GIRF) by relying on Koop et al. (1996). For a vector Z_t the Generalized Impulse Response defines the reaction of Z_t to a shock δ_j on ϵ_{jt}, conditional on the information set (I_{t-1}) at time $t-1$ as

$$GI_z\left(n, \delta_j, I_{t-1}\right) = E\left(Z_{t+n} | \epsilon_{jt} = \delta_j, I_{t-1}\right) - E\left(Z_{t+n} | I_{t-1}\right)$$

This formula doesn't rely on an orthogonalization procedure (e.g Cholesky), and the interpretation is straight-forward. In fact instead of shocking all the system, only the jth variable is shocked and the effect of the other variables are integrated out.

If Z_t is d-dimensional vector having the following Moving Average representation

$$Z_t = \sum_{i=1}^{\infty} A_i \varepsilon_{t-i}$$

with ε_t has a normal distribution, the integration is easily done using the formula

$$E\left(\varepsilon_t | \varepsilon_{jt} = \delta_j\right) = \left(\sigma_{1j}, \sigma_{2j}, \ldots, \sigma_{dj}\right)' \sigma_{jj}^{-1} \delta_j = \Omega e_j \sigma_{jj}^{-1} \delta_j$$

where e_j is the vector having 1 at the jth position and 0 elsewhere.

The vector of the unscaled impulse response of the effect of a shock in the jth equation at time t on Z_{t+n} is given by

$$GI_z\left(n\right) = A_n \delta_j = A_n \Omega e_j \sigma_{jj}^{-1} \delta_j \ , n = 0,\ 1,\ 2,\ ...$$

Then normalizing the size of the shock to a one standard deviation $\delta_j = \sqrt{\sigma_{jj}}$ gives

$$GI_z\left(n\right) = \sigma_{jj}^{-\frac{1}{2}} A_n \Omega e_j$$

2.3.2 Definition of the measure

Relying on the intuition of the generalized impulse response, we propose to compute the square of the permanent component response to a shock in each markets. This may also be interpreted as a forecast variance error, but we are not interested here in its decomposition.

let's write the long run impact matrix as $\Psi(1) = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$ and design its two rows by $\psi_1 = \begin{pmatrix} \psi_{11} & \psi_{12} \end{pmatrix}$ and $\psi_2 = \begin{pmatrix} \psi_{21} & \psi_{22} \end{pmatrix}$.

With the cointegrating vector $\beta = (1 \ -\beta_1)$ and the properties $\beta'\Psi(1) = 0$ we have

$$\begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} = \begin{pmatrix} \psi_{11} - \beta_1\psi_{12} \\ \psi_{12} - \beta_1\psi_{22} \end{pmatrix} = 0$$

Thus $\psi_{11} = \beta_1\psi_{12}$ and $\psi_{12} = \beta_1\psi_{22}$, which implies that the second row is multiple of the first row $\psi_2 = \beta_1^{-1}\psi_1$. There is 1 cointegrating relation so the space of permanent component is of dimension 1 and $\Psi(1) = \begin{pmatrix} 1 \\ \beta_1^{-1} \end{pmatrix} \psi_1$.

The permanent component entering the first price p_{1t} is given by $Q_{1t} = \psi_1 \sum_{s=1}^{t} \varepsilon_s$ it is a random walk, the same identified by the information share when $\beta_1 = 1$. To define our measure we compute the generalized impulse response of Q_{1t}, to a shock in the first and the second market. The response of Q_{1t} after n periods to a shock to the jth market is given by

$$\begin{aligned} GI_{Q_1}(n, \varepsilon_j) &= E(Q_{1t+n}|\varepsilon_{jt} = \delta_j, \Omega_{t-1}) - E(Q_{1t+n}, \Omega_{t-1}) \\ &= \psi_1 E(\varepsilon_t|\varepsilon_{jt} = \delta_j) \\ &= \psi_1 \Omega e_j \sigma_{jj}^{-\frac{1}{2}} \\ &= \psi_1 \sigma_{jj}^{-\frac{1}{2}} (\sigma_{1j}, \sigma_{2j})' \\ &= \sigma_{jj}^{-\frac{1}{2}} (\psi_{11}\sigma_{1j} + \psi_{12}\sigma_{2j}) \end{aligned}$$

The horizon n disappears from the formula at the second equality thanks to the random walk nature of Q_{1t}.

The square of the impulse response gives the variance of the permanent component forecast error resulting from the shock in the jth market. As the IS using the variance, it is a good summary of the permanent information entering the prices by market j:

$$GI_{q_1}(\varepsilon_j)^2 = \sigma_{jj}^{-1} (\psi_{11}\sigma_{1j} + \psi_{12}\sigma_{2j})^2$$

As it doesn't not sum-up to one, the contribution of the jth market to price discovery that we called Impulse Response Information Share $IRIS$ is defined by

$$IRIS_j = \frac{\sigma_{jj}^{-1}\left(\psi_{11}\sigma_{1j} + \psi_{12}\sigma_{2j}\right)^2}{\sum_{l=1}^{2}\sigma_{ll}^{-1}\left(\sum_{i=1}^{d}\psi_{1i}\sigma_{il}\right)^2} \tag{9}$$

Remember that the $IRIS$ was computed using Q_{1t} the permanent component entering the first price. If we consider the permanent component entering the second market: $Q_{2t} = \psi_2 \sum_{s=1}^{t}\varepsilon_s = \beta_1^{-1}Q_{1t}$. It is a multiple of Q_{1t} so the impulse response of Q_{2t} to a shock to the jth price is:

$$GI_{Q_2}(\varepsilon_j) \quad = E\left(Q_{2t+n}|\varepsilon_{jt} = \delta_j, \Omega_{t-1}\right) - E\left(Q_{2t+n}, \Omega_{t-1}\right) = \beta_1^{-1} \quad GI_{Q_1}(\varepsilon_j)$$

and

$$IRIS_j = \frac{\beta_1^{-1}GI_{Q_2}(\varepsilon_j)^2}{\sum_{j=1}^{2}\beta_1^{-1}GI_{Q_2}(\varepsilon_j)^2} = \frac{GI_{Q_1}(\varepsilon_j)^2}{\sum_{j=1}^{2}GI_{Q_1}(\varepsilon_j)^2}$$

So $IRIS$ doesn't depend on which permanent component you choose[4] and the estimation of the VECM can be done without imposing the unit restriction on the cointegrating equation.

2.3.3 Relationship between $IRIS$ and the Information Share measures.

To present the link between $IRIS$ and the IS of Hasbrouck (1995), we write explicitly the formulas for market 1.

$$IRIS_1 = \frac{\sigma_{11}^{-1}\left(\psi_{11}\sigma_{11} + \psi_{12}\sigma_{12}\right)^2}{\sigma_{11}^{-1}\left(\psi_{11}\sigma_{11} + \psi_{12}\sigma_{12}\right)^2 + \sigma_{22}^{-1}\left(\psi_{11}\sigma_{12} + \psi_{12}\sigma_{22}\right)^2}$$

If Ω is diagonal then the $IRIS$ measure gives the IS measure:

$$IRIS_1 = \frac{\sigma_{11}^{-1}\psi_{11}^2\sigma_{11}^2}{\sigma_{11}^{-1}\psi_{11}^2\sigma_{11}^2 + \sigma_{22}^{-1}\psi_{12}^2\sigma_{22}^2} = \frac{\psi_{11}^2\sigma_{11}}{\psi_{11}^2\sigma_{11} + \psi_{12}^2\sigma_{22}}$$

If Ω is not diagonal, let $\sigma_{12} = \rho\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$ and let's consider the expression of the Cholesky roots of Ω when the market 1 is placed first in vector of prices (Baillie et al., 2002):

$$F = \begin{pmatrix} \sqrt{\sigma_{11}} & 0 \\ \rho\sqrt{\sigma_{22}} & \sqrt{\sigma_{22}}\sqrt{(1-\rho^2)} \end{pmatrix}$$

[4]This properties is easily seen for more than two markets, there is $d-1$ cointegration relations so only 1 permanent component entering each market multiplied by the corresponding β_i^{-1}

Then the numerator of the Information Share of the market 1 which correspond to Hasbrouck (1995) upper bound is

$$\begin{aligned}([\psi F]_1)^2 &= \left(\psi_{11}\sqrt{\sigma_{11}} + \psi_{12}\rho\sqrt{\sigma_{22}}\right)^2 \\ &= \psi_{11}^2\sigma_{11} + \psi_{12}^2\rho^2\sigma_{22} + 2\psi_{11}\psi_{12}\rho\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\end{aligned}$$

let's now focus on the numerator of $IRIS_1$, the square of the Impulse Response of the permanent component to a shock in market 1.

$$\begin{aligned}\sigma_{11}^{-1}\left(\psi_{11}\sigma_{11} + \psi_{12}\sigma_{12}\right)^2 &= \sigma_{11}^{-1}\left(\psi_{11}\sigma_{11} + \psi_{12}\rho\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\right)^2 \\ &= \psi_{11}^2\sigma_{11} + \psi_{12}^2\rho^2\sigma_{22} + 2\psi_{11}\psi_{12}\rho\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}\end{aligned}$$

So the numerator of the $IRIS$ for the first market corresponds to the numerator of the IS for the first market when it is in the first position in the orthogonalization procedure. We can thus write

$$IRIS_1 = \frac{Upper.IS1 \times \psi\Omega\psi'}{Upper.IS1 \times \psi\Omega\psi' + Upper.IS2 \times \psi\Omega\psi'} = \frac{Upper.IS1}{Upper.IS1 + Upper.IS2}$$

This formula strongly supports the use of Hasbrouck (1995) upper bound to solve the identification problem of IS. Many studies choose randomly between the lower-bound and the upper bound, or the mid-point for their application. But those choices are done without a clear justification, our framework thus provide an economics sensitive rationale for the use of the upper-bound Information Share.

2.3.4 Estimation and Testing

The computation of the impulse response is easy once the parameters of the VECM representation of the price have been identified. With real data after selecting the order K with the help of information criteria, the VECM (33) is estimated for $\hat{\Omega}$ and for the parameters $\Gamma_1, \ldots, \Gamma_K$. Then $\hat{\Psi}(1)$ is computed and the elements of the impulse response are identified.

To obtain standard errors and the limiting distribution of the response to jth market, we make use of the limiting distribution of the coefficients. we have

$$\hat{GI}_j = \hat{\psi}_1\hat{\Omega}e_j\hat{\sigma}_{jj}^{-\frac{1}{2}}$$

For the deduction of asymptotic the VECM (33) is be represented as

$$\Delta Y = -\alpha\beta' Y_{-1} + \mathbf{\Gamma}\Delta X + U$$

with T= sample size and
$\Delta Y \coloneqq [\Delta P_1, \dots \Delta P_T]$
$Y_{-1} \coloneqq [P_0, \dots P_{T-1}]$
$\mathbf{\Gamma} \coloneqq [\Gamma_1, \dots, \Gamma_K]$
$U \coloneqq [\epsilon_0, \dots \epsilon_T]$
$\Delta X \coloneqq [\Delta X_0, \dots \Delta X_T]$ with $\Delta X_{t-1} \coloneqq \begin{bmatrix} \Delta P_{t-1} \\ \vdots \\ \Delta P_{t-K} \end{bmatrix}$

The theorem 1 gives the asymptotic distribution for the response of the permanent component to a shock in the jth market .

Theorem 1. *Let P_t the vector of prices satisfying VECM (33)). then*

$$\sqrt{T}\left(\hat{GI}_j - GI_j\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{i\hat{r}is}\right)$$

With

$$\Sigma_{iris} = \sigma_{jj}^{-1}\left(e_j'\Omega \otimes e_j'\right) F \Sigma_\gamma F' \left(\Omega e_j \otimes e_j\right) + \sigma_{jj}^{-1}\left(e_j'\Psi(1) \otimes e_j'\right) \Sigma_{\hat{\sigma}} \left(\Psi'(1) e_j \otimes e_j\right)$$

Where the notations are defined in the following:
$\Sigma_{\hat{\sigma}} = 2D_K \left(D_K' D_K\right)^{-1} D_K' \left(\Omega \otimes \Omega\right)$
$\Sigma_\gamma = plim\, T \begin{bmatrix} \beta' Y_{-1} Y_{-1}' \beta & \beta' Y_{-1} \Delta X' \\ \Delta X Y_{-1}' \beta & \Delta X \Delta X' \end{bmatrix}^{-1} \otimes \Omega$
$F = \left(\left(\Psi'(1)\left(I - \sum_{i=1}^K \Gamma_i\right)' - I_d\right) H_{\alpha_1'} \left(\alpha' H_{\alpha_1'}\right)^{-1}, \left(\iota_K' \otimes \Psi'(1)\right)\right) \otimes \Psi'(1)$
D_K *(duplication matrix) and* $H_{\alpha_1'}$ *are matrix of 0-1 (defined in appendix).* ι_K *represents a column vector with ones of length* K

The asymptotic variance can be used to computed standard errors and test the significance of the permanent response to a shock in one market. For this purpose, the different expressions in the variance are simply replaced by their feasible estimator. The matrix $H_{\alpha_1'}$ might be bit a tricky to built especially if the cointegration rank is not known. Fortunately in our setup the rank of cointegration is known to be $d-1$ and provided that α' is put in reduced echelon form we have the $[d \times (d-1)]$ matrix

$$H_{\alpha_1'} = \begin{bmatrix} 0_{1\times d-1} \\ I_{d-1} \end{bmatrix}$$

For example in the bivariate case we have $H_{\alpha_1'} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

To test the significance of GI_j the following testing statistic is computed

$$\hat{ST}_j = \sqrt{T}\frac{|\hat{GI}_j|}{\hat{\Sigma}_{iris}}$$

and can be compared with the critical values of the classical t-test.

3 Price discovery measures in Continuous time

Up to our knowledge there is no literature on price discovery measures in continuous time. Meanwhile today financial data are available in frequency that are more and more close to continuous time. The use of high frequency data come with some news features that are extensively studied in volatility estimation literature. The core point being the consideration of what is called microstructure noise. The present framework aims to provide the literature with a measure of price discovery that is suitable to high frequency data features. We will name it High Frequency Information Share (HFIS) to stay on the same terminology. The importance of designing a measure for high frequency data comes principally from the distortion in the variance estimation when noises are present.

Example 1. Consider one asset that is continuously traded on two markets 1 and 2 at respective prices p_{1t} and p_{2t}. The fundamental log price of the asset is $dm_t = \sigma dW_t$ with W_t standard Brownian motion. The data are recorded at discrete points and the observed prices $p_t = (p_{1t}, p_{2t})'$ correspond to m_t plus some i.i.d non correlated noises $(\varepsilon_{1t}, \varepsilon_{2t})$

$$\begin{aligned} p_{1t} =& m_t + \varepsilon_{1t} \\ P_{2t} =& m_t + \varepsilon_{2t} \end{aligned}$$

Let h be the discretization pace, we have

$$\Delta p_{t+h} = \begin{bmatrix} p_{1t+h} - p_{1t+h} = & m_{t+h} - m_t + \varepsilon_{1t+h} - \varepsilon_{1t} \\ p_{2t+h} - p_{1t+h} = & m_{t+h} - m_t + \varepsilon_{2t+h} - \varepsilon_{2t} \end{bmatrix} = \begin{bmatrix} \sqrt{h}\sigma N(0,1) & + & \Delta\varepsilon_{1t+h} \\ \sqrt{h}\sigma N(0,1) & + & \Delta\varepsilon_{2t+h} \end{bmatrix}$$

then

$$Var\left(\Delta p_{t+h}\right) = \begin{bmatrix} h\sigma + 2\sigma^2_{\varepsilon_1} & h\sigma \\ h\sigma & h\sigma + 2\sigma^2_{\varepsilon 2_1} \end{bmatrix} \xrightarrow{h\to 0} \begin{bmatrix} 2\sigma^2_{\varepsilon_1} & 0 \\ 0 & 2\sigma^2_{\varepsilon 2_1} \end{bmatrix}$$

When h goes to zero the variance of the price difference is driven only by microstructure noises variances. So at high frequency in the presence of microstructure noises, measures of price discovery presented above are not really based on the efficient price variance but rather on non-informative noise.

3.1 Setup

There is one asset that is traded on two markets 1 and 2. The true log price of the asset is Y_t, this is the price which in finance literature will result in perfect world , with no arbitrage.

We define by $X_t = (X_{1t}, X_{2t})$ the price we would have observed on market 1 and 2 if there were not non-informative noise (This correspond to observed prices in Hasbrouck framework). This may be formalized by saying that $X_{1t} = Y_t + \mu_{1t}$ and $X_{2t} = Y_t + \mu_{2t}$

$\mu_t = (\mu_{1t}, \mu_{2t})$ is the part of microstructure noise that is completely related to information, this is for example due to asymmetric information. Its is consistent with what Menkveld et al. (2007) calls market's over-reaction (or under-reaction) to information[5].

The model is set in continuous time but the data are observed at the discrete points $(t_i)_{i=1,..,n} \in [0, T]$ corresponding to the different instants of trade. The recorded prices $P_t = (P_{1t}, P_{2t})'$ are X_t plus a contamination.

We have the following observation equations:

$$\begin{aligned} P_{1th} &= X_{1th} + \varepsilon_{1th} \\ P_{2th} &= X_{2th} + \varepsilon_{2th} \end{aligned} \tag{10}$$

$\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})$is the part of microstructure noise not related to information (tick size, measurement error...).

In summary this is to say that the observed prices are equal to the efficient price plus one information-related noise and, one non-informative noise. This is already a difference with the discrete time setup where the noise is driven by the same innovation as the permanent component such that we have only informative noise. We make the following assumptions

[5]This is consistent with Menkveld(2007) where he estimates a price equation comprising the sum of three elements. The first term is the efficient price, the second term is the market over/under-reaction to information proportionnal to the efficient price innovation, and the third part is the microstructure noise arising from bid ask spread and price discreteness.

Assumption 1. *Conditionally on process X the noise ε_{t_i} is i.i.d with $E(\varepsilon_{t_i}/X) = 0$.*

Assumption 2. *X_{1t}, X_{2t} and X_t are non-stationary but their increments are covariance stationary.*

Since all errors are assume stationary this is the same as saying that the true price increment or the observe price increment are stationary.

Assumption 3. *The spread between the two prices $X_{1t} - X_{2t}$ is covariance stationary*

It is reasonably thought that if prices are far one from another, traders can made profit by buying on one market and selling on another. So arbitrage and fast connection between the different market places strengthen this argument. Meanwhile it rules out deterministic time varying volatility.

The direct implication of these assumptions is that the prices are cointegrated and we will thus use the analog of the tools in in discrete time.

Notations

The length of a trading period is denoted by T (for example one day of observations), it is the period on which our contribution is going to be computed. Ω_t is use for a volatility matrix and $\Omega_t^2 \equiv \Omega_t \Omega_t'$ for a the Covariance matrix

We recall here some useful notations and operators to write ARMA class processes in continuous time, this is presented extensively in Cochrane (2012).

Lag operator: $L^s X_t = X_{t-s}$ $s \in \mathbb{R}$

Difference operator $DX_t = \frac{1}{dt} dX_t$

The link between the two is : $D = -logL$ and $L = e^{-D}$. In continuous time, unlike discrete time, working with D rather than L renders the formula tractable and eases the computation.

Those operators permit to write operator function that parallel the $MA(\infty)$ representation in discrete time. For example the $MA(\infty)$ representation of a gaussian stationary process y_t can be written as:

$$y_t = \int_{\tau=0}^{\infty} c(\tau)\sigma dB_{t-\tau} = \mathcal{L}_c(D)\Omega DB_t \tag{11}$$

The term $\Omega dB_{t-\tau}$ represents the innovation, and the Laplace transform operator function ($\mathcal{L}$) is

$$\mathcal{L}_c(D) = \int_{\tau=0^-}^{\infty} c(\tau)e^{-D\tau}d\tau$$

Using these notations we develop the construction of the HFIS measure of price discovery.

3.2 Construction of the measure

3.2.1 Identification of the permanent component

In continuous time under assumptions 1 to 3, the Wold's representation theorem argument:

$$DX_t = \mathcal{L}_c(D)DW_t = \int_{\tau=0^-}^{\infty} c(\tau)e^{-D\tau}dW_t \tag{12}$$

where W_t is a process of orthogonal increments and we assume the following:

Assumption 4. *$DW_t = \Omega_t DB_t$ where is a standard Brownian motion. Ω_t is a zero-mean stationary stochastic volatility with $E(\Omega_t^2) = \Omega^2$*

The vector X_t is cointegrated with cointegrating vectors $\beta = \begin{pmatrix} 1 & -\lambda_1 \end{pmatrix}$,
Using the Beveridge-Nelson decomposition of the difference operator function

$$\mathcal{L}_c(D) = \mathcal{L}_c(0) + D\mathcal{L}_b(D) \tag{13}$$

This parallel the discrete time version where $D\mathcal{L}_b(D)$ is the proper level operator function of a stationary process.

then replacing in 12 we get

$$DX_t = \mathcal{L}_c(0)\Omega_t DB_t + D\mathcal{L}_b(D)\Omega_t DB_t$$

This allow a decomposition in term of common stochastic trend and stationary component. This decomposition display the fact that

$$DX_t = DZ_t + Dw_t \tag{14}$$

where $w_t = \mathcal{L}_b(D)DB_t$ is a stationary process, and $Z_t = X_t - w_t$ is a martingale (even a pure random walk) satisfying

$$DZ_t = \mathcal{L}_c(0)\Omega_t DB_t \tag{15}$$

We restate the difference with IS where X_t is the observed process and then the transitory part of price w_t has its innovations correlated with the innovation in the martingale component. here X_t is unobserved and will be dirtied by the other sources of microstructure noise.

If $\mathcal{L}_c(0) \neq 0$ and is not full rank $r < n$ then X_t is cointegrated, and there exist a matrix

β' such that $\beta' X_t$ is stationary. There is also another matrix α summarizing the impact of the long common trend on the variable X_t.

Cointegration required $\beta' \mathcal{L}_c(0) = 0$, and thus the row of $\mathcal{L}_c(0)$ are collinear. They are the same when the cointegrating coefficient is equal to 1. If this is not the case the second row $\psi_{2r} = \lambda_1^{-1} \psi_{1r}$ so using one or the other will not change the value of our propose measure. Let us take the first row and define $\psi = \psi_{1r}$ then we have

$$\mathcal{L}_c(0) = J\psi \, , with \, J = (1 \; \lambda_1^{-1})^\top$$

Now from the formula 15 of the martingale part we have

$$dZ_t = J\psi \Omega_t dB_t \tag{16}$$

where we extract the common component to all the markets which is taken as the fundamental price of the asset.

$$z_t = z_0 + \psi \int_0^t \Omega_s dB_s \tag{17}$$

3.2.2 The Impulse response function

Consider the permanent component of prices given by equation 17. In the spirit of IRIS we define the impulse response of this permanent price to a shock in market 1. For this we rewrite the permanent prices in term of original shocks W_t and consider a shock of value of a $\delta_{1t} = dW_{1t}$, so the response of the dz_t to a shock in market 1 is

$$GI_1(t) = E\left(dz_t | dW_{1t} = \delta_{1t}\right) \tag{18}$$

This measure how the permanent price reacts to a news in the first market. This is a good intuitive property to define a measure of price discovery, in the sense that fundamental information impacts permanently the prices.

3.2.3 Constant Volatility case

We start our development by assuming that the volatility matrix Ω is constant, which allows a comprehensive presentation of our construction. Using the normality properties of the

Brownian motion and the conditional expectation formula we derive

$$\begin{aligned} GI_1(t) &= \psi E\left(dW_t | dW_{1t} = \delta_{1t}\right) \\ &= \psi \Omega e_1 \sigma_{11}^{-1} \times \delta_{1t} \\ &= \psi \left(\begin{array}{cc} \sigma_{11}, & \sigma_{12} \end{array} \right)' \sigma_{11}^{-1} \times \delta_{1t} \end{aligned}$$

Where . Thus the cumulative response on a period $[0, T]$ is given by

$$\begin{aligned} CGI_1(t) &= \psi \int_0^T \left(\begin{array}{cc} \sigma_{11}, & \sigma_{12} \end{array} \right)' \sigma_{11}^{-1} \times \delta_{1t} \\ &= \psi \int_0^T \left(\begin{array}{cc} \sigma_{11}, & \sigma_{12} \end{array} \right)' \sigma_{11}^{-1} dW_{1t} \end{aligned} \tag{19}$$

As Hasbrouck (1995) using the variance, the quadratic variation of this cumulative response is good summary of the amount of information due to market 1:

$$\begin{aligned} < CGI_1 > &= \left[\psi \left(\begin{array}{cc} \sigma_{11}, & \sigma_{12} \end{array} \right)' \sigma_{11}^{-1} \right]^2 \times < dW_{1t} > \\ &= \left[\psi \left(\begin{array}{cc} \sigma_{11}, & \sigma_{12} \end{array} \right)' \sigma_{11}^{-1} \right]^2 \times \sigma_{11} T \end{aligned} \tag{20}$$

Similarly for a shock on the second market we obtain that the cumulative response variation of the permanent component is

$$< CGI_2 > = \left[\psi \left(\begin{array}{cc} \sigma_{12}, & \sigma_{22} \end{array} \right)' \sigma_{22}^{-1} \right]^2 \times \sigma_{22} T \tag{21}$$

Thus we define the *"High Frequency Information Share"* (HFIS) measure of price discovery as

$$HFIS_1 = \frac{< CGI_1 >}{< CGI_1 > + < CGI_2 >}$$

If we replace the value it gives.

$$HFIS_1 \equiv \frac{(T\sigma_{11})^{-1} \left[\psi_{11}(T\sigma_{11}) + \psi_{12}(T\sigma_{12})\right]^2}{(T\sigma_{11})^{-1} \left[\psi_{11}(T\sigma_{11}) + \psi_{12}(T\sigma_{12})\right]^2 + (T\sigma_{22})^{-1} \left[\psi_{11}(T\sigma_{12}) + \psi_{12}(T\sigma_{22})\right]^2} \tag{22}$$

We add T in front of each the volatility coefficient to keep visible the continuous-time feature of the framework, this will be important for the presentation of the next section and to

understand the estimation strategy.

3.2.4 The stochastic volatility case

When we relax the assumption of a constant volatility, the spot variance matrix is now dependent on t and in previous framework the generalized response of the permanent component is

$$GI_1(t) \quad = \quad \psi \left(\ \sigma_{11t}, \quad \sigma_{12t} \ \right)' \sigma_{11t}^{-1} \times \delta_{1t} \tag{23}$$

Let's denote by $\bar{\sigma}_{ijt} = \int_0^T \sigma_{ijt} dt$ and $\bar{\Omega}^2 = \int_0^T \Omega_t^2 dt$. By mimicking the previous construction we propose the following formula for our $HFIS$.

$$HFIS_1 = \frac{\psi \bar{\sigma}_{11t}^{-1} \left[\psi_{11}\bar{\sigma}_{11t} + \psi_{12}\bar{\sigma}_{12t}\right]^2}{\bar{\sigma}_{11t}^{-1} \left[\psi_{11}\bar{\sigma}_{11t} + \psi_{12}\bar{\sigma}_{12t}\right]^2 + \bar{\sigma}_{22t}^{-1} \left[\psi_{11}\bar{\sigma}_{12t} + \psi_{12}\bar{\sigma}_{22t}\right]^2} \tag{24}$$

For d markets with $d \geq 2$, the contribution of market j is

$$HFIS_j = \frac{\left[\psi \bar{\Omega}^2 e_j\right]^2 \bar{\sigma}_{jjt}^{-1}}{\sum_{j=1}^{d} \left[\psi \bar{\Omega}^2 e_j\right]^2 \bar{\sigma}_{jjt}^{-1}} \tag{25}$$

Where e_j is the vector having 1 at the j^{th} position and 0 elsewhere.

Some comments need to be done on this formula. In fact as there is the inverse of the variance in the formula of CGI_1, the formula 24 is not what appears exactly when computing the quadratic variation. In fact applying strictly the quadratic variation of the cumulative process of GI_1 in equation 23 will give a formula where the coefficient $\int_0^T \sigma_{12t} / \int_0^T \sigma_{11t}$ is replaced by $\int_0^T \left(\sigma_{12t}/\sigma_{11t}\right) dt$. This is not really an issue as first it is a definition and it doesn't have an impact on our relative measure. Second it will complicate the presentation and makes the estimation unfeasible without additional assumptions. Assuming for example that the volatility parameters are constant piecewise allows, an estimation of this ratio can be done block-by-block using the methods presented in section 3.

Another important feature of our framework is that it provides a generalization to continuous time of the IS of Hasbrouck (1995). In fact the total information entering z_t on $[0, T]$ can be represented by its quadratic variation:

$$S_T = \frac{1}{T} \psi \left(\int_0^T \Omega_t^2 dt \right) \psi'$$

and by taking F as the Cholesky root of $\left(\int_0^T \Omega_t^2 dt\right)$, the contribution of market j defined

by formula 7 is:

$$IS_j = \left([\psi F]_j\right)^2 / S_T \tag{26}$$

and by switching ordering one gets lower and upper bound on information share.

4 Estimation

The estimation of the measure is done by computing each element of the formula 24 which include components of the integrated covariance matrix ($\int_0^T \Omega_t^2 dt$) and elements of the vector ψ. The volatility parameters will be estimated using existing covolatility estimators in recent econometrics of High frequency data and ψ is estimated through the weighting matrix of the classical VECM equation.

4.1 Basic Case

We make the following assumption that we will relax latter. We will assume that the $MA(\infty)$ representation represents a continuous time AR(1) process.

As can be seen, though this is a simplifying assumption, it is general enough in continuous time finance literature.

Assumption 5. *X_t admits the following continuous time AR representation*

$$DX_t = -\Pi X_t + \Omega_t DB_t$$

4.1.1 Estimation of ψ

With cointegration properties the assumption 5 corresponds to the ECM

$$dX_t = -\alpha\beta' X_t dt + \Omega_t dB_t \tag{27}$$

For the estimation a discretisation scheme should be made. Advantages of the exact discretisation form of process are highlighted by Phillips (1991), Comte (1999), Chambers (1999, 2011). The exact discrete form of this Equation 27 is given by

Proposition 1. *Using the exact discrete form of equation 27, the following VECM representation applies to P_t*

$$\Delta P_{th} = g(h,d)\,\alpha\beta' P_{th-h} + \xi_{th} \tag{28}$$

where $\xi_{th} = e_{th} + \Delta u_{th} + g(h,d)\,\alpha\beta' u_{th-h}$,

$e_{th} = \int_{th-h}^{th} e^{-(th-s)\alpha\beta'} \Omega_s dB_{th-s}$
and $g(h,d) = 1 - e^{-hd}$

Using the representation 28,$\hat{\alpha}$ is estimated consistently (see proposition 2), $\mathcal{L}_c(0)$ is computed using the formula

$$\mathcal{L}_c(0) = I - \alpha(\beta'\alpha)^{-1}\beta'$$

and then ψ is identified as a row of $\mathcal{L}_c(0)$.

Proposition 2. *Let*

- $Z_{th} = \beta' P_{th} = P_{1th} - P_{2th}$
- $\hat{\alpha} = h^{-1}\left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\Delta P_{th}\right)$ [6]

when $Th \to \infty$ *and* $Th \times h^2 \to 0$

$$\sqrt{Th}\,(\hat{\alpha} - \alpha) \Longrightarrow O(1)\sqrt{2\Omega_u} W(1)$$

Where W *is the standard wiener process*

Proof. See Appendix □

4.1.2 Estimation of $\int_0^T \Omega_t^2 dt$

From Assumptions 1 to 5, using the observed noisy values, the system to estimate is:

$$\begin{cases} dX_t &= \mu_t dt + \Omega_t dB_t \\ P_t &= X_t + \varepsilon_t \end{cases} \tag{29}$$

With $\mu_t = \Pi X_t$

The integrated volatility $\int_0^T \Omega_t^2 dt$ of the process 29 is estimated in the recent high frequency econometrics framework. We propose with proposition 3 to use the Modulated Realised Covariance (MRC) estimator of Christensen et al. (2010) which is robust to microstucture noise. We use the suboptimal estimator which is always definite positive. It is minor issue here since the second part of the estimator is to remove the asymptotic bias coming entirely from noise.

[6] This is like and IV estimator of α in equation 28 using Z_{th-2h} as Instrument to solve for endogeneity due to measurements errors

Proposition 3. *In equation 34,*

$$MRC[Y]_n^\delta = \frac{1}{\psi_2 k_n^\delta} \sum_{i=0}^{n-k_n+1} \bar{Y}_n^i (\bar{Y}_n^i)' \tag{30}$$

$$MRC[Y]_n^\delta \xrightarrow{h\to 0} \int_0^T \Omega_t^2 dt \text{ with } h: \text{ discretization pace}$$

where the different notations are presented in appendix.

In practice in this framework, the drift is known after the estimation of α. It can strongly influences the results if we are far from the assumption (mainly the boundedness) of the preaveraging method. So we remove the drift first and then computed the preaveraged return with the demeaned series.

4.2 Generalization

We will present how the previous framework is adapted to more general setting.

Estimation of ψ

According to the Granger representation theorem of a cointegrated time series, there exist a representation in vector error correction form:

$$\mathcal{L}_d(D)DX_t = -\alpha\beta' X_t + \Omega_t DB_t \tag{31}$$

$$\int_{\tau=0}^{\infty} d(\tau) dX_{t-\tau} = -\alpha\beta' X_t + \Omega_t dB_t$$

The assumption 5 corresponds to the case where $\mathcal{L}_d(D) = I$.

Discretizing this equation in the simple scheme to estimate is α,

$$\Delta X_t = -\alpha\beta' X_{t-1} + \Gamma_1 \Delta X_{t-1} + \ldots + \Gamma_K \Delta X_{t-K} + e_t \tag{32}$$

The last lag K can be chosen by information criteria.

With the observed value P_t we obtain the following linear regression model with autocorrelated error

$$\Delta P_t = -\alpha\beta' P_{t-1} + \Gamma_1 \Delta P_{t-1} + \ldots + \Gamma_K \Delta P_{t-K} + \xi_t \tag{33}$$

α (thus ψ) can be again estimated by linear regression, IV can be use with past value of $\beta' P_t$ as instruments. Consistency is difficult to established formally in this case when

$T \to \infty$ and $h \to 0$. And especially because K should be moving, but as in the discrete time literature this is the same problem as choosing the K. And by simulation we check that we have consistency when $Th \to \infty$ and $h \to 0$.

Estimation of $\int_0^T \Omega_t^2 dt$

From equation 12, to impose a restriction on the first term, it is assume a Dirac Delta in $c(\tau)$ at $\tau = 0$ such that its Laplace transform is $c_0 = 1$. that is the contemporaneous impact of the noise of the price is one,

$$DX_t = c(0)\Omega_t DB_t + \int_{\tau=0}^{\infty} c(\tau)e^{-D\tau}\Omega_t DB_t$$

$$dX_t = \left(\int_{\tau=0}^{\infty} c(\tau)\Omega_t dB_{t-\tau}\right) dt + c(0)\Omega_t dB_t = \mu_t dt + \Omega_t dB_t \tag{34}$$

So replacing in 10 the system to estimate is

$$\begin{cases} dX_t &= \mu_t dt + \Omega_t dB_t \\ P_t &= X_t + \varepsilon_t \end{cases} \tag{35}$$

Which is estimated using the MRC.

4.3 Simulation

Two markets with private information

Market 1 drives the efficient price, and market 2 price relies on the lagged value of m_t. Clearly all price discovery happens in market 1

$\mu_{1t} = \sigma_{u1} dW_{1t}$
$\mu_{2t} = \sigma_{u2} dW_{2t}$
$m_t = \int_0^t \mu_{1s} = \sigma_{u1} W_{1t}$
with W_{1t} and W_{2t} standard Brownian motion

$$\begin{cases} p_{1t} &= m_t + \mu_{1t} + \varepsilon_{1t} \\ p_{2t} &= m_{t-\delta} + \mu_{2t} + \varepsilon_{2t} \end{cases}$$

- $\varepsilon_{1t} \sim \mathcal{N}(0, \sigma_1^2)$ and $\varepsilon_{2t} \sim \mathcal{N}(0, \sigma_1^2)$, $E(\varepsilon_{1t}\varepsilon_{2t}) = 0$
- $\sigma_{1m} = \sigma_{2m} = 1$, $\sigma_1 = \sigma_2 = 0.0005$,

[Insert Table 1 here]

HFIS is the only one that assign almost perfectly dominance to market 1. It is the only one that detect the true dominance when the time lag become very small like 1second.

5 Empirical application

We will use our measure of daily price discovery to study the relative part US exchanges in the price discovery of some asset comprising the SP500 and are traded on NYSE, NASDAQ and regional exchanges.

The data comes from the TAQ Database and will be done on the year 2011. We will work at a frequency of 1 second. Before the application a number of cleaning action have been done on the raw data:

First we suppress the data stamped before the opening (9h00) and after the closing (17h30) of the market will also remove the data between 9h05 because the activity at the opening session create a phenomenon in the data that can not be related to the purpose of this study. Second to handle the asynchronicity problem, we will fill the data with the last trade price.

5.1 Descriptive analysis

Market analysis:

NYSE and NASDAQ are the two biggest exchanges in the world by capitalization and trade value. NYSE remains by far the first with a capitalization of around 14 USD trillion in 2011 (around 16 USD trillion in 2014). During this year the trade value was about 20 USD trillions, so an average daily amount of 55 USD billions.

NASDAQ has a market capitalization of 4.6 USD trillions, and a trade value of 13.5 USD trillions, corresponding to and average value of an average daily amount of 37 USD billions. (http://www.i3investor.com/jsp/hti/usmarket.jsp). Beside these two giant exchanges another one is very important is the FINRA (Financial industry and regulatory authority) it is in the database but the trades reported there seems to be mostly hidden trades ("Dark pool"), so no price discovery is suppose to happen there.

[Insert Table 2 here]

For the places where specific assets are traded the domination is not that pronounced as shown by the average daily statistics in table. For Apple, FINRA concentrate almost

1/3 of the trading volume, and NYSE almost 17% when NASDAQ traded around 1/3 also, the remaining being traded principally by DirectEdge and BATS. NASDAQ is the principal market for Tech. stocks, but on our study period with 15% of traded shares, NASDAQ is dominated by NYSE (28%) and FINRA (29 %); the same pattern is observed for HEWLETT PACKARD, Microsoft and American Express. Concerning Cisco most of the shares are exchanged on FINRA (38%).

[Insert Table 3 here]

In term of liquidity (we think of liquidity here as the frequency of transactions) , a look at the number of transaction per day shows that NASDAQ cumulates the biggest number of transaction for apple (32%) and American express (27%). This is not in contradiction with the previous results on volumes; it simply tells us that most big-size transactions are done on FINRA and NYSE.

The second group of asset with study is concern with asset that trades a volume of the order of 5 million and is less liquid than the first at the order of 50 000 transactions per day. For that asset there most of the shares exchanged are via FINRA (30% for GNW, 32 % for LLY and 29% for Nike). In term of liquidity FINRA shows 1/3 of the number of transactions, which is around one trade per second, for GNW and LLY while NYSE is the most liquid for NE (35%).

The third group comprised asset with around 1 USD million and are traded less than 10 000 times in a day. For those assets NYSE appear to concentrate most of the volume and the number of transaction for all the assets. For AVERY DENNISON, 30.4% (on NYSE in New-york) and 13.7% (on NYSE ARCA in Chicago) of shares are traded, while 15% are on the NASDAQ and 20% on FINRA. This is the same order as the percentages for the number of transactions, even if the exchanges recorded less than 1-2 thousands transactions per day.

The repartition of the bound on daily information of ABX price show, that most of the price discovery happens in NYSE.

5.2 Results on markets contribution

The results on markets contributions in table 4 shows that for Apple most of the price discovery happens on the NYSE even if NASDAQ is the primary exchange of APPLE. It is followed by NASDAQ and FINRA is also important with almost the same weight as APPLE. This result confirm and obvious fact that the segmentation and the liberalization really changed the structure of the market. For comparison in Hasbrouck (1995) study for

almost all asset NYSE concentrated most of the trade and in its results almost 90 % of the contribution to price movement.

in summary: Price discovery happens generally on NYSE (for big and medium size trade) and NASDAQ (for small size trade), the contribution of a market is correlated with its market share for small and medium size transactions (table 5), and it is not large quantity trade that convey information. This is consistent with the thought that informed traders split their orders in small quantities trade.

[Insert Table 5 here]

6 Conclusion

This papers aim at providing the literature with a new framework for assessing price discovery. The literature on this topic have mainly used a discrete time representation of the price dynamic with some drawback highlighted in the previous sections. First, the paper proposes to study the impact of a shock in one market on the permanent component of prices by using the Generalized impulse response function providing an invariant measure of price discovery with economic relevance. In a continuous time setup we propose the High-Frequency Information share (HFIS) measure of price discovery that extend the idea that and is not affected by difference in level of non-informative microstructure noise.

We apply the methodology to a bunch of SP500 stocks traded on some notable US exchanges. We found that price discovery happens generally on NYSE (for big and medium size trade) and NASDAQ (for small size trade), the contribution of a market is correlated with its market share for small and medium size transactions, and it is not large quantity trade that convey information. This is consistent with the thought that informed traders split their orders in small quantities trade. This measure of price discovery can then be used on empirical analysis to investigate what are the driving channels of information for cross-listed asset, for example in a regression where there is HFIS on the left-hand side and on the right-hand side variables like Number of participant, latency, liquidity, trade size, location...

References

Andersen, T., T. Bollerslev, F. Diebold, and P. Labys (2000). Great realizations. *RISK 13*, 105–108. 3

Baillie, R. T., G. G. Booth, Y. Tse, and T. Zabotina (2002). Price discovery and common factor models. *Journal of Financial Markets 5*(3), 309 – 321. 3, 9, 12

Barclay, M. J., T. Hendershott, and D. T. M. N (2003). Competition among trading venues: Information and trading on electronic communications networks. *Journal of Finance 58*, 2637–2665. 4

Blume, M. E. and M. A. Goldstein (1997). Quotes, order flow, and price discovery. *The Journal of Finance 52*(1), 221–244. 4

Chakravarty, S., H. Gulen, and S. Mayhew (2004). Informed trading in stock and option markets. *The Journal of Finance 59*(3), 1235–1257. 4

Chambers, M. J. (1999). Discrete time representation of stationary and non-stationary continuous time systems. *Journal of Economic Dynamics and Control 23*(4), 619 – 639. 22

Chambers, M. J. (2011). Cointegration and sampling frequency. *The Econometrics Journal 14*(2), 156–185. 22

Christensen, K., S. Kinnebrock, and M. Podolskij (2010). Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data. *Journal of econometrics 159*(1), 116–133. 23

Cochrane, J. H. (2012). Continuous-time linear models. *Foundations and TrendsÂ® in Finance 6*(3), 165–219. 17

Comte, F. (1999). Discrete and continuous time cointegration. *Journal of Econometrics 88*(2), 207 – 226. 22

De Jong, F. (2002). Measures of contributions to price discovery: a comparison. *Journal of Financial Markets 5*(3), 323 – 327. 3

Eun, C. S. and S. Sabherwal (2003a). Cross-border listings and price discovery: Evidence from u.s.-listed canadian stocks. *The Journal of Finance 58*(2), 549–575. 3

Eun, C. S. and S. Sabherwal (2003b). Cross-border listings and price discovery: Evidence from u.s.-listed canadian stocks. *The Journal of Finance 58*(2), 549–575. 9

Gonzalo, J. and C. Granger (1995). Estimation of common long-memory components in cointegrated systems. *Journal of Business & Economic Statistics 13*(1), 27–35. 3, 8, 9

Grammig, J. and F. J. Peter (2013). Telltale tails: A new approach to estimating unique market information shares. *Journal of Financial and Quantitative Analysis 48*, 459–488. 4, 8

Harris, F. H., T. H. McInish, and R. A. Wood (2002a). Common factor components versus information shares: a reply. *Journal of Financial Markets 5*(3), 341 – 348. 9

Harris, F. H., T. H. McInish, and R. A. Wood (2002b). Security price adjustment across exchanges: an investigation of common factor components for dow stocks. *Journal of Financial Markets 5*(3), 277 – 308. 3

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance 50*(4), 1175–1199. 3, 4, 6, 7, 9, 10, 12, 13, 20, 27

Hasbrouck, J. (2002). Stalking the efficient price in market microstructure specifications: an overview. *Journal of Financial Markets 5*(3), 329 – 339. 3

Huang, R. D. (2002). The quality of ecn and nasdaq market maker quotes. *The Journal of Finance 57*(3), 1285–1319. 4

Jacod, J., Y. Li, P. A. Mykland, M. Podolskij, and M. Vetter (2009). Microstructure noise in the continuous case: The pre-averaging approach. *Stochastic Processes and their Applications 119*(7), 2249 – 2276. 37

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in gaussian vector autoregressive models. *Econometrica 59*(6), 1551–1580. 5

Koop, G., M. Pesaran, and S. Potter (1996). Impulse response analysis in nonlinear multivariate models. *Journal of Econometrics 74*(1), 119–147. 10

Lehmann, B. N. (2002). Some desiderata for the measurement of price discovery across markets. *Journal of Financial Markets 5*(3), 259 – 276. 3

Lien, D. and K. Shrestha (2009). A new information share measure. *Journal of Futures Markets 29*(4), 377–395. 8

Lien, D. and K. Shrestha (2014). Price discovery in interrelated markets. *Journal of Futures Markets 34*(3), 203–219. 3, 4, 7

Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated. 32

Menkveld, A. J., S. J. Koopman, and A. Lucas (2007). Modeling around-the-clock price discovery for cross-listed stocks using state space methods. *Journal of Business & Economic Statistics 25*(2), 213–225. 4, 16

Pesaran, H. and Y. Shin (1998). Generalized impulse response analysis in linear multivariate models. *Economics Letters 58*(1), 17 – 29. 4, 10, 32, 33

Phillips, P. (1991). Error correction and long-run equilibrium in continuous time. *Econometrica 59*(4), 967–980. 22

Vlaar, P. J. G. (2004). On the asymptotic distribution of impulse response functions with long-run restrictions. *Econometric Theory 20*(5), 891–903. 32

Yan, B. and E. Zivot (2010). A structural analysis of price discovery measures. *Journal of Financial Markets 13*(1), 1 – 19. 3, 9

A Appendix

A.1 Proofs

A.1.1 Proof of theorem 1

Proof. As it is shown in Lütkepohl (2007) the asymptotic is the same considering that β is known the reason being that $\hat{\beta}$ is estimated at the rate T better than the $\sqrt{T}$ of $\hat{\alpha}$. To simplify the formulas we use Ψ to denote $\Psi(1)$.

Let $\gamma \coloneqq vec[\alpha : \mathbf{\Gamma}]$,

The *vec* operator (stacking the column of matrix) and the *vech* operator that stacks the elements on and below the diagonal of a square matrix.

The following asymptotic and the expression for Σ_γ and $\Sigma_{\hat{\sigma}}$ are derived from Lütkepohl (2007) and Pesaran and Shin (1998) :

- $\sqrt{T}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(0, \Sigma_\gamma)$
- $\sqrt{T} vec\left(\hat{\Omega} - \Omega\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_{\hat{\sigma}})$
- The duplication matrix D_K is the $\left(K^2 \times \frac{1}{2}K(K+1)\right)$ matrix of 0-1 such that for any $(K \times K)$ matrix A, $vec(A) = D_K vech(A)$.
- The estimators of $[\alpha : \mathbf{\Gamma}]$ and Ω are asymptotically independent.

As Ψ depends only on the $[\alpha : \mathbf{\Gamma}]$ we have

$$\sqrt{T}\begin{bmatrix} vec\left(\hat{\Psi} - \Psi\right) \\ vec\left(\hat{\Omega} - \Omega\right) \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{bmatrix} \Sigma_\psi & 0 \\ 0 & \Sigma_{\hat{\sigma}} \end{bmatrix}\right) \tag{36}$$

To obtain the asymptotic variance of Ψ as a function of γ the Delta-method gives

$$\Sigma_\psi = \frac{\partial vec\Psi(1)}{\partial \boldsymbol{\gamma}'} \Sigma_\gamma \frac{\partial vec\Psi(1)'}{\partial \boldsymbol{\gamma}} \tag{37}$$

and the expression for $\frac{\partial vec\Psi(1)}{\partial \boldsymbol{\gamma}'}$ is derived from Vlaar (2004):

$$F = \frac{\partial vec\Psi(1)}{\partial \boldsymbol{\gamma}'} = \left(\left(\Psi'\left(I - \sum_{i=1}^{K}\Gamma_i\right)' - I_d\right) H_{\alpha_1'}\left(\alpha' H_{\alpha_1'}\right)^{-1}, \left(\iota_K' \otimes \Psi'\right)\right) \otimes \Psi' \tag{38}$$

The matrix $H_{\alpha_1'}$ is a matrix of 0-1 selecting the column of α' such that $\left(\alpha' H_{\alpha_1'}\right)$ is non singular and $H_{\alpha_2'}$ selects the remaining columns. Those matrix allows the representation

$\alpha_{\perp} = H_{\alpha_2'} - H_{\alpha_1'}\left(\alpha' H_{\alpha_1'}\right)^{-1}\alpha' H_{\alpha_2'}$.
The response of the market j is given by

$$GI_j = \psi_1 \Omega e_j \sigma_{jj}^{-\frac{1}{2}} = \left(e_j' \Omega e_j\right)^{-\frac{1}{2}} \times e_j' \Psi \Omega e_j$$

it estimator is thus

$$\begin{aligned}
\hat{GI}_j &= \hat{\psi}_1 \hat{\Omega} e_j \hat{\sigma}_{jj}^{-\frac{1}{2}} \\
&= e_j' \hat{\Psi} \hat{\Omega} e_j \times \left(e_j' \hat{\Omega} e_j\right)^{-\frac{1}{2}} \\
&= \left(e_j' \left(\hat{\Psi} - \Psi\right)\left(\hat{\Omega} - \Omega\right) e_j + e_j' \left(\hat{\Psi} - \Psi\right) \Omega e_j + e_j' \Psi \left(\hat{\Omega} - \Omega\right) e_j + e_j' \Psi \Omega e_j\right) \times \left(e_j' \hat{\Omega} e_j\right)^{-\frac{1}{2}} \\
&= Ng/Dg
\end{aligned}$$

For the denominator we have from Pesaran and Shin (1998) that given the consistency of the ML estimators there is a scalar R which is $o_p(1)$ such that

$$Dg = \left(e_j' \Omega e_j\right)^{\frac{1}{2}} + \frac{1}{2}\left(e_j' \Omega e_j\right)^{\frac{1}{2}} \left(e_j' \otimes e_j'\right) vec\left(\hat{\Omega} - \Omega\right) + R = \sigma_{jj}^{\frac{1}{2}} + o_p(1) \tag{39}$$

We now compute the distribution of the numerator

$$\begin{aligned}
\sqrt{T}\left(Ng - e_j' \Psi \Omega e_j\right) &= e_j' \left(\hat{\Psi} - \Psi\right)\left(\hat{\Omega} - \Omega\right) e_j + e_j' \left(\hat{\Psi} - \Psi\right) \Omega e_j + e_j' \Psi \left(\hat{\Omega} - \Omega\right) e_j \\
&= \left(e_j' \Omega \otimes e_j'\right) \sqrt{T} vec\left(\hat{\Psi} - \Psi\right) + \left(e_j' \Psi \otimes e_j'\right) \sqrt{T} vec\left(\hat{\Omega} - \Omega\right) + o_p(1) \qquad (40) \\
&= \left[\left(e_j' \Omega \otimes e_j'\right),\ \left(e_j' \Psi \otimes e_j'\right)\right] \begin{bmatrix} \sqrt{T} vec\left(\hat{\Psi} - \Psi\right) \\ \sqrt{T} vec\left(\hat{\Omega} - \Omega\right) \end{bmatrix} + o_p(1) \qquad (41)
\end{aligned}$$

The second equality uses the following straightforward relations

$$\begin{cases}
e_j' \left(\hat{\Psi} - \Psi\right) \Omega e_j & = \left(e_j' \Omega \otimes e_j'\right) vec\left(\hat{\Psi} - \Psi\right) \\
e_j' \Psi \left(\hat{\Omega} - \Omega\right) e_j & = \left(e_j' \Psi \otimes e_j'\right) vec\left(\hat{\Omega} - \Omega\right) \\
\left(\hat{\Psi} - \Psi\right)\left(\hat{\Omega} - \Omega\right) & = o_p\left(1/\sqrt{T}\right)
\end{cases}$$

From the results in formulas 39, 41and equation **??**, $\sqrt{T}\left(\hat{GI}_j - GI_j\right)$ is asymptotically normal with variance

$$\begin{aligned}\Sigma_{iris} &= \sigma_{jj}^{-1}\left[\left(e_j^{'}\Omega\otimes e_j^{'}\right),\ \left(e_j^{'}\Psi\otimes e_j^{'}\right)\right]\begin{bmatrix}\Sigma_{\psi} & 0\\ 0 & \Sigma_{\hat{\sigma}}\end{bmatrix}\left[\left(e_j^{'}\Omega\otimes e_j^{'}\right),\ \left(e_j^{'}\Psi\otimes e_j^{'}\right)\right]^T\\ &= \sigma_{jj}^{-1}\left(e_j^{'}\Omega\otimes e_j^{'}\right)\Sigma_{\psi}\left(\Omega e_j\otimes e_j\right)+\sigma_{jj}^{-1}\left(e_j^{'}\Psi\otimes e_j^{'}\right)\Sigma_{\hat{\sigma}}\left(\Psi^{'}e_j\otimes e_j\right)\end{aligned}$$

□

A.1.2 Proof of proposition 1

Proof. By solving the differential equation represented by 27 for a given h

$$\begin{aligned}X_{th} &= e^{-h\alpha\beta}X_{th-h}+e_{th}\\ \Delta X_{th} &= \left(e^{-h\alpha\beta}-I\right)X_{th-h}+e_{th}\end{aligned}\tag{42}$$

whith $e_{th}=\int_{th-h}^{th}e^{-(th-s)\alpha\beta^{'}}\Omega_s dB_{th-s}$.
$exp\left(-h\alpha\beta^{'}\right)=\sum_{l=0}^{\infty}\left(-h\right)^l\left(\alpha\beta^{'}\right)^l$
$\left(\alpha\beta^{'}\right)^2=\alpha\beta^{'}\alpha\beta^{'}=d\times\alpha\beta^{'}$ where $d=\beta^{'}\alpha=\alpha_1-\alpha_2$
by recurrence $\left(\alpha\beta^{'}\right)^l=d^{l-1}\times\alpha\beta^{'}$ and we have

$$\begin{aligned}exp\left(-h\alpha\beta^{'}\right)-I &= \left(-h\alpha\beta^{'}+\left(-h\right)^2\left(\alpha\beta^{'}\right)^2/2+\dots\right)\\ &= \alpha\beta^{'}\left(-h+\left(-h\right)^2d/2+\dots\right)\\ &= \tfrac{\alpha\beta^{'}}{d}\left(-hd+\left(-h\right)^2d^2/2+\dots\right)\\ &= \tfrac{\alpha\beta^{'}}{d}\left(-1+1-hd+\left(-h\right)^2d^2/2+\dots\right)\\ &= \alpha\beta^{'}\tfrac{\left(-1+e^{-hd}\right)}{d}\\ &= -\alpha\beta^{'}g(h,d)\end{aligned}$$

With $g(h,d)=\frac{\left(1-e^{-hd}\right)}{d}$,
replacing in the expression 42 gives

$$\Delta X_{th}=-g\left(h,d\right)\alpha\beta^{'}X_{th-h}+e_{th}$$

With the observed value P_t :

$$\begin{aligned}\Delta P_{th} &= \Delta X_{th}+\Delta u_{th}\\ &= -g(h,d)\alpha\beta^{'}X_{th-h}+\Delta u_{th}+e_{th}\\ &= -g(h,d)\alpha\beta^{'}P_{th-h}-g(h,d)\alpha\beta^{'}u_{th-h}+\Delta u_{th}+e_{th}\\ &= -g(h,d)\alpha\beta^{'}P_{th-h}+\xi_{th}\end{aligned}$$

with $\xi_{th} = e_{th} + \Delta u_{th} - g(h,d)\alpha\beta' u_{th-h}$ □

Lemma 1.

1. $g(h,d) = h + O(h^2)$
2. $Var(e_{th}) = h\Omega^2 + O(h^2)$
3. $Var(\xi_{th}) = 2\Omega_u^2 + O(h)$
4. $E\left(\xi_{th}\xi_{th-h}'\right) = \Omega_u^2 + g(h,d)\alpha\beta'\Omega_u^2 = \Omega_u^2 + O(h)$

Proof. .

1. $g(h,d) = \frac{\left(1-e^{-hd}\right)}{d} = h - (-h)^2 d/2 + \ldots = h + O(h^2)$
2. $Var(e_{th}) = Var\left(\int_{th-h}^{th} e^{-(th-s)\alpha\beta'}\Omega_s dB_s\right) = Var\left(\int_0^h e^{-u\alpha\beta'}\Omega_s dB_u\right)$
   $Var(e_{th}) = \int_0^h (I - g(h,u)\alpha\beta')\Omega^2(I - g(h,u)\alpha\beta')' du$
   $Var(e_{th}) = h\Omega^2 - \left(\int_0^h g(h,u)du\right)\alpha\beta'\Omega^2 - \left(\int_0^h g(h,u)du\right)\Omega^2\beta\alpha' + \int_0^h g(h,u)^2\alpha\beta'\Omega^2\beta\alpha' du$
   Using 1) we have the result: $Var(e_{th}) = h\Omega^2 + O(h^2)$
3. $\xi_{th} = (e_{th} + u_{th} - (1 - g(h,d)\alpha\beta')u_{th-h})$

$$\begin{aligned} Var(\xi_{th}) &= Var(e_{th}) + \Omega_u^2 + (1 - g(h,d)\alpha\beta')\Omega_u^2(1 - g(h,d)\beta\alpha') \\ &= h\Omega^2 + O(h^2) + 2\Omega_u^2 - g(h,d)\left(\alpha\beta'\Omega_u^2\right) - g(h,d)\left(\Omega_u^2\beta\alpha'\right) + g(h,d)^2\left(\alpha\beta'\Omega_u^2\beta\alpha'\right) \\ &= 2\Omega_u^2 + \left(\Omega^2 - \alpha\beta'\Omega_u^2 - \Omega_u^2\beta\alpha'\right)O(h) + O(h^2) \end{aligned}$$

4. .

$$\begin{aligned} E\left(\xi_{th}\xi_{th-h}'\right) &= E\left((e_{th} + u_{th} - (1 - g(h,d)\alpha\beta')u_{th-h})(e_{th-h} + u_{th-h} - (1 - g(h,d)\alpha\beta')u_{th-2h})'\right) \\ &= (1 - g(h,d)\alpha\beta')\Omega_u^2 \\ &= \Omega_u^2 - g(h,d)\alpha\beta'\Omega_u^2 \end{aligned}$$

□

Lemma 2. *Let* $Z_{th} = \beta' P_{th} = P_{1th} - P_{2th}$ *then*
$Var(Z_{th}) = O(h^{-1})$
$E(Z_{th}Z_{th-h}) = O(h^{-1})$

Proof. $Z_{th} = \beta' P_{th} = P_{1th} - P_{2th}$

$$\begin{aligned}
\Delta P_{th} &= g(h,d)\alpha Z_{th-h} + \xi_{th} \\
\Delta \beta' P_{th} &= g(h,d)\beta'\alpha Z_{th-h} + \beta'\xi_{th} \\
\Delta Z_{th} &= g(h,d)d \times Z_{th-h} + \beta'\xi_{th} \\
Z_{th} &= (1 + g(h,d)d) \times Z_{th-h} + \beta'\xi_{th} \\
Z_{th} &= e^{-hd} \times Z_{th-h} + \beta'\xi_{th}
\end{aligned}$$

$$\begin{aligned}
Var(Z_{th})\left(1 - e^{-2hd}\right) &= Var(\beta'\xi_{th}) + Cov\left(\beta'\xi_{th} Z'_{th-h}\right) \\
&= \beta' Var(\xi_{th})\beta + \beta' Cov\left(\xi_{th}\xi'_{th-h}\right)\beta \\
&= \beta'\left(2\Omega_u^2 + O(h)\right)\beta + \beta'\left(\Omega_u^2 + O(h)\right)\beta \\
Var(Z_{th})\left(-2hd + O\left(h^2\right)\right) &= 3\beta'\Omega_u^2\beta + O(h) \\
Var(Z_{th}) &= \frac{3\beta'\Omega_u^2\beta + O(h)}{h(-2d + O(h))} \\
&= O\left(h^{-1}\right)
\end{aligned}$$

$$\begin{aligned}
Cov\left(Z_{th}, Z'_{th-h}\right) &= e^{-hd} Var(Z_{th-h}) + Cov\left(\beta'\xi_{th} Z'_{th-h}\right)\beta' Var(\xi_{th})\beta + \beta' Cov\left(\xi_{th}\xi'_{th-h}\right)\beta \\
&= e^{-hd} O\left(h^{-1}\right) + \beta'\left(\Omega_u^2 + g(h,d)\alpha\beta'\Omega_u^2\right)\beta\beta'\left(2\Omega_u^2 + O(h)\right)\beta + \beta'\left(\Omega_u^2 + O(h)\right)\beta \\
&= e^{-hd} O\left(h^{-1}\right) + \beta'\Omega_u^2\beta + O(h) \\
&= O\left(h^{-1}\right)
\end{aligned}$$

$$\begin{aligned}
E\left(Z_{th-2h}\xi'_{th}\right)^2 &= Var(Z_{th-h}) \times Var(\xi_{th}) \\
&= O\left(h^{-1}\right) \times \left(2\Omega_u^2 + O(h)\right)
\end{aligned}$$

□

A.1.3 Proof of proposition 2

Now we prove the proposition using the Lemma 1 to replace the variable by the bigs O results

$$\begin{aligned}
\hat{\alpha} &= h^{-1}\left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\Delta P_{th}\right) \\
h\hat{\alpha} &= \left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\Delta P_{th}\right) \\
&= -g(h,d)\alpha + \left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\xi_{th}\right)
\end{aligned}$$

$$\begin{aligned}
h\hat{\alpha} - h\alpha &= (-g(h,d) + h)\alpha + \left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\Delta P_{th}\right) \\
h(\hat{\alpha} - \alpha) &= O\left(h^2\right) + \left(\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\left(\sum_{t=1}^{T} Z_{th-2h}\xi_{th}\right) \\
\sqrt{Th}(\hat{\alpha} - \alpha) &= O\left(\sqrt{T}h^2\right) + \left(T^{-1}\sum_{t=1}^{T} Z_{th-1h} Z_{th-2h}\right)^{-1}\frac{1}{\sqrt{Th}}\left(\sum_{t=1}^{T} Z_{th-2h}\xi_{th}\right) \\
&\approx O\left(\sqrt{T}h^2\right) + \left(Cov\left(Z_{th}, Z'_{th-h}\right)\right)^{-1}\frac{1}{\sqrt{h}}\sqrt{Var(Z_{th-2h})\,Var(\xi_{th})} \times W(1) \\
&\approx O\left(\sqrt{T}h^2\right) + \left(O\left(h^{-1}\right)\right)^{-1}\sqrt{O(h)}\frac{1}{\sqrt{h}}\sqrt{2\Omega_u^2 + O(h)} \times W(1) \\
&\approx O\left(\sqrt{Th}h\right) + O(1) \times \sqrt{2\Omega_u^2 + O(h)} \times W(1) \\
&\Longrightarrow O(1)\sqrt{2\Omega_u^2} \times W(1)
\end{aligned}$$

Provided that $\sqrt{Th} \to \infty$ and $\sqrt{Th} \times h \to 0$.

A.1.4 Notations of proposition 3:

The notations in 3 originally come from the pre-averaging method of Jacod et al. (2009), which provides and estimator of the integrated volatility,

Assuming that the true log price is generated by an Îto process of the form

$$X_t = X_0 + \int_0^t u_s ds + \int_0^t \sigma_s dW s$$

where W is a standard wiener process and $\mu = (\mu_t)$ and $\sigma = (\sigma_t)$ are adapted processes.

and the noisy observed process is given by

$$Z_t = X_t + \epsilon_t$$

let k_n be the size of each group in the length of each period at the first stage. it is chosen such that

$k_n = \theta * n^{0.5+\delta} + o(n^{\frac{1}{4}})$

We have a function g on $[0,1]$, continuous, piecewise C^1 with a piecewise Lipschitz derivative g'satisfying $g(0) = g(1) = 0$, and $\int_0^1 g(x)^2 dx > 0$. let's define some quantities and notation.

$g_i^n = g(i/k_n)$,

$\phi_1(s) = \int_s^1 g'(u)g'(u-s)du, \qquad \phi_2(s) = \int_s^1 g(u)g(u-s)du,$

$for\, s > 1, \phi_1(s) = 0, \quad \phi_2(s) = 0,$

$\phi_{ij} = \int_0^1 \phi_i(s)\phi_j(s)ds, \qquad \psi_i = \phi_i(0),\ i,j = 1,2$

their empirical equivalent:

$\widehat{\psi}_1 = k_n \sum_{j=1}^{k_n}(g_{j+1}^n - g_j^n)^2 \qquad \widehat{\psi_2}(j) = \frac{1}{k_n}\sum_{j=1}^{k_n-1}(g_j^n)^2\widehat{\phi}_1(j) = \sum_{i=j+1}^{k_n}(g_{j-1}^n - g_i^n)(g_{i-j-1}^n - g_{i-j}^n) \qquad \widehat{\phi}_2(j) = \sum_{i=j+1}^{k_n} g_i^n g_{j-1}^n$

$\widehat{\phi}_{11}(j) = k_n\left(\sum_{j=0}^{k_n-1}(\widehat{\phi}_1(j))^2 - \frac{1}{2}(\widehat{\phi}_1(0))^2\right) \quad \widehat{\phi}_{12}(j) = \frac{1}{k_n}\left(\sum_{j=0}^{k_n-1}\widehat{\phi}_1(j)\widehat{\phi}_2(j) - \frac{1}{2}\widehat{\phi}_1(0)\widehat{\phi}_2(0)\right)$

Then, pre-averaged return are defined as:

$$Z_i^n = Z_{i\Delta_n}, \quad \Delta_i^n Z = Z_i^n - Z_{i-1}^n, \quad \overline{Z}_i^n = \sum_{j=1}^{k_n-1} g_i^n \Delta_{i+j}^n Z$$

B TABLES

Table 1: Two markets with private information

Lag δ	HFIS	IRIS	ISu	ISl	CS
0s	99.17	10.76	60.1	44.0	22.4
	(2.8)	(10.6)	(2.3)	(2.3)	(9.6)
1s	96.9	11.4	57.7	42.3	23.6
	(3.6)	(11.1)	(2.4	(2.4	(10.3)
2s	97.0	11.8	56.7	43.3	23.6
	(2.8)	(14.5)	(1.8)	(1.8)	(11.7)
3s	96.7	17.2	56.7	43.3	28.1
	(3.2)	(12.7)	(2.5)	(2.5)	(10.9)
4s	95.8	14.0	57.4	42.6	25.5
	3.2	11.6	2.3	2.3	10.3

Table 2: *Average daily number and volume of transactions, on the different markets from the 01/03 to 30/05/2011*. liq=number of transactions per day; vol=volume of trade per day (in 10^3)

AAPL			
	Vol	share	liq
FINRA	4 464	32	18 818
NASDAQ	4 023	29	27 614
NYSE ARCA	2 350	17	17 349

AXP			
	Vol	share	liq
FINRA	1 574	24	6 065
NYSE	1 505	23	5 724
NASDAQ	1 340	21	9 222
NYSE ARCA	675	10	4 547

UTX			
	Vol	share	liq
NYSE	923	26	14 434
FINRA	788	22	12 313
NASDAQ	770	21	12 035
NYSE ARCA	452	13	7063

AVY			
	Vol	share	liq
NYSE	236	30	1 679
FINRA	162	21	1 519
NASDAQ	121	16	1 034
NYSE ARCA	106	14	1 201

Table 3: *Market share of transactions by transactions size(in %)*

AAPL		
size	0-180	>180
FINRA	15	37
NASDAQ	35	29
NYSE ARCA	24	17

AXP		
size	0-160	>160
FINRA	12	19
NYSE	12	33
NASDAQ	29	18
NYSE ARCA	21	12

UTX		
size	0-160	>160
NYSE	15	37
FINRA	14	20
NASDAQ	24	14
NYSE ARCA	25	15

AVY		
size	0-130	>130
NYSE	19	**45**
FINRA	14	16
NASDAQ	**22**	16
NYSE ARCA	18	12

Table 4: *Contribution of each market to price discovery for the assets differents assets. The table report the High frequency information share computed in a 5 dimensional VAR*

	AXP	AVY	UTX	AAPL
FINRA	6.43	3.8	6.43	6.41
NYSE	**42.38**	**50.6**	**38.8**	
NYSE ARCA	13.19	18.2	12.38	**26.94**
NASDAQ	**31.81**	18.6	**34.49**	**46.10**
BATS	6.17	8.59	8.90	10.13
DIRECTX				10.3

Table 5: *Correlation of the contribution to price discovery with volume market share for different transactions size.*

		HFIS	share all	share <150	share 150-200	share >200
NYSE	share all	0.612	1			
	share <150	0.394	0.91	1		
	share 150-200	**0.715**	0.73	0.45	1	
	share >200	0.688	0.46	0.13	0.46	1
NASDAQ	share all	0.74	1			
	share <150	**0.73**	0.93	1		
	share 150-200	0.315	0.58	0.37	1	
	share >200	0.008	0.14	-0.16	0.46	1
FINRA	share all	0.024	1			
	share <150	0.028	0.97	1		
	share 150-200	0	0.87	0.79	1	
	share >200	-0.024	0.36	0.22	0.42	1