Mispricing, Learning, and Price Discovery

We extend the "information share" (IS) framework of De Jong and Schotman (2010) by accommodating the endogenous error-correction mechanism proposed by Andersen et al. (2022) and co-movement pricing-errors across markets. We discuss identification issues and propose a generalized measure of information shares. We apply the new framework to Sp500-Emini and SPY ETF contracts and find evidence that our model can capture the intraday patterns of IS and other market behaviors more accurately. Specifically, our model reveals the intraday patterns of single market self-learning (endogenous error-correction mechanism), multiple markets cross-learning (cointegration error-correction mechanism), over/under-reaction to the efficient price shock, and the IS of SPY and SP500-EMINI markets.

1 Introduction

Learning is an important feature in many microstructure models. Some traders have private information and they trade on it; other traders observe market data and they learn from it (O'Hara, 2015).

In any market, there are various information sources from which to learn. One natural source of learning is the historical price of the asset itself (regardless of other markets). Imperfections in the information set and learning conclusions compounded by acquisition and processing delays (Andersen et al., 2022) cause price movements that do not fully reflect the information innovation or are dilute by noise. When investors learn this, prices adjust accordingly and the endogenous pricing-errors[1] are corrected. This endogenous pattern is vital from two perspectives: (1) Theoretically, the endogenous error-correction mechanism answers a long standing puzzle – the dynamics of autocorrelation. For the past two decades, researchers have pointed out that autocorrelation behavior can vary a lot across different assets and periods, especially since the sign of autocorrelation can dramatically change. However, this pattern is still not accounted for in most related studies, which could be questionable. According to Hansen and Lunde (2006), exogenous noise representations are inconsistent with high-frequency asset return dynamics, as exhibited by the autocorrelation pattern. (2) Empirically, for models with only exogenous noise representations, in which traders do not learn from the past information of the underlying asset's historical price, or in which endogenous 'look-back' learning is not incorporated, such price-errors diverge from the efficient price and the spot price in the long run and are hardly corrected. Nevertheless, the endogenous pricing-error pattern is underestimated in previous literature. One potential reason might be the existence of identification issues. For the models embedded with endogeneity learning patterns (e.g., endogenous pricing-errors), the estimation and identification of the parameters can be cumbersome. Moreover, state-space models (unobserved component models), argued by various literature as one of the most efficient and powerful frameworks for high-frequency market puzzles, commonly suffer from strong identification issues due to the latent variable(s). Thus, the endogenous pricing-error or error-correction models within the state-space framework, which could have a very high potential to capture

[1]The difference between the historical spot price and historical efficient value.

various market behaviors, are not well widely adopted since the estimation of these models typically requires strong additional assumptions, which could be hard to support economically.

Only recently, there have been several milestone exercises. Andersen et al. (2022) introduce their model with an endogenous error-correction mechanism to allow for autocorrelation sign and return dynamics changes. They find a robust stylized fact that indicates the existence of a significant endogenous error-correction mechanism within the market. In other words, many investors are learning from past pricing errors to correct current prices. However, because of the natural difficulty of estimating state-space representations, they have to make stronger assumptions to solve the identification issue led by the endogenous pricing-error terms – they assume that ongoing learning on the part of the investor moves slower, at a constant frequency, than the efficient price innovation and endogenous error-correction terms. Here, we extend their endogenous error-correction model to the multi-market case which relaxes the need for identification constraints and generates more stable estimation.

Another natural source of learning is the price dynamics of related markets. Therefore, our framework is also designed to capture the cross-learning pattern among multiple markets. Similarly, when investors learn pricing-errors from closely-linked markets, they take advantage of this pricing-error so that it will be eliminated/minimized by the underlying error-correction mechanism. Such a pattern is essential, especially recently, since markets are more tightly inter-connected, sewn together by market making/statistical arbitrage that operates across, not just within, markets (O'Hara, 2015). Specifically, some of these markets are naturally connected by arbitrage or short-term equilibrium considerations. It is natural to wonder how the pricing-errors (e.g., the price disagreement between decentralized markets) are corrected in the price discovery process or the error-correction mechanism of multiple markets. For example, closely-linked markets where prices have a cointegrating relationship (i.e., the same asset on different exchanges). It is critical to examine cointegration type pricing-errors (the price difference of these market prices) and the underlying error-correction mechanism because traders constantly seek arbitrage opportunities by the price difference of cointegrated series. Hence, our framework is also designed to capture such pricing-errors and the underlying error-correction mechanism.

Furthermore, for closely linked multiple market cases, one essential question of price discovery is 'who moves first from the efficient price innovations.' One of the most popular measurements of price formation is the information share (IS) of Hasbrouck (1995), which is based on the cointegration model of Engle and Granger (1987). Other approaches have also been advocated: Harris, McInish, Shoesmith, and Wood (1995) propose the Component Share (CS) framework; Yan and Zivot (2010) and Putniņš (2013) merge IS and CS into new indicators, and De Jong and Schotman (2010) extend the IS to an unobserved component (UC) framework. However, none of these approaches is ideal. The VECM/VMA IS frameworks can be easily diluted by the nature of Cholesky decomposition, and the UC framework requires strong assumptions to solve the identification issue. Thus, we provide a new approach and framework to solve these problems.

In general, motivated by the studies of endogenous error-correction mechanism by Andersen et al. (2022) and the IS estimation with state-space representation by De Jong and Schotman (2010), we propose a novel solution to the issues above: we propose a framework that extends both the framework of Andersen et al. (2022) and the framework of De Jong and Schotman (2010). Our model naturally connects multivariate frameworks in Hasbrouck (1995) and De Jong and Schotman (2010) to two critical pricing-errors and underlying error-correction mechanisms. Hence, the benefits are three-fold: (1) In comparison to the framework of Andersen et al. (2022), we prove the identification issue is solved within our framework, and our framework better captures the patterns of the multivariate markets because we include further information (the closely-linked market relationship). (2) Compared to other multivariate frameworks, our framework not only solves the identification issue of these frameworks, but also further captures the intraday pattern of two pricing-errors, cointegration and endogenous pricing errors, and estimates the underlying error-correction mechanisms. Since the learning patterns have been more critical in the past decade, we believe our framework is more realistic by including the two error-correction mechanisms as the close-linked markets' further co-movement mechanisms. (3) By elaborating the IS, our framework can capture it more accurately. Moreover, our model captures the intraday pattern of IS, rarely discussed in previous literature.

In summary, our contributions are: (1) We propose a novel solution to the identification issue of the Andersen et al. (2022) endogenous pricing-error model. (2) We extended the De Jong and Schotman (2010) state-space IS framework by allowing two additional error-correction mechanisms. (3) Our work fills the gap between various market microstructure models and the IS framework. It is hard for researchers to calculate IS from most market microstructure models under previous IS frameworks, but many of these models readily fit into our generalized framework which enables the calculation of the IS. (4) Our model provides a feasible and better methodology to estimate the intraday pattern of the market, e.g., the scale IS, error-correction mechanisms, and over/under-reaction to efficient price shocks, without requiring ultra-high frequency data. With tick level data, our framework can generate intraday IS per 10 minutes. Specifically, we find i. the endogenous pricing-error is corrected faster in the SP500-EMINI market; ii. the cross-learning pattern between these two markets is stronger shortly after the opening and before the closing of the SPY market; iii. the two markets have similar IS, but they are not consistent on intraday level. The SPY is more sensitive to efficient price shock shortly after the opening of the SPY market, and the EMINI is more sensitive shortly after the opening and before the closing of the SPY market. (5) Our model further connects the market microstructure models with high-frequency econometrics.

The rest of the paper is structured as follows. Section 2 introduces the model setup, including elaboration of the identification issue and the IS. Section 3 focuses on the discussion of the model estimation and empirical exercise. Section 4 presents the empirical findings, and Section 5 concludes the paper.

2 Model Set Up

2.1 General framework

In our multivariate framework, we allow for generalized error-correction terms that can include various kinds of pricing-error and underlying error-correction mechanisms (endogenous error-correction type terms[2] and cointegration error-

[2]The difference of the past spot price and the efficient price.

correction type terms[3]), and the spot value of these markets shares the same efficient price (from the permanent-transitory decomposition). Therefore, our general model is written as follows:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \Phi_L(p_t, m_t) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{1}$$

Where m_t is the scalar of efficient price at time t, r_t is the scalar of the efficient price change at time t. For the general K cointegrated markets case, p_t is a K $\times$ 1 vector containing the spot price of K markets, α is a K $\times$ 1 vector, $\vec{1}$ is a K $\times$ 1 vector of 1, Φ_L is a matrix of functions of lagged linear combinations of p_t and m_t. e_t is a K $\times$ 1 vector of normally distributed noise terms. In our framework, we restrict $Cov(r_t, r_{t-i}) = 0$, $Cov(e_t, e_{t-i}) = \{0\}$, $Cov(r_i, e_j) = \vec{0}$ for all i, j $\in \mathbb{N}$, where $\{0\}$ is a K $\times$ K matrix of 0 and $\vec{0}$ is a K $\times$ 1 vector of 0.

To fit the cointegration relationships of each pair of markets within these K markets, we decompose $\Phi_L(p_t, m_t)$ as $\Phi_L(p_t, m_t) = \Phi_{1L}(p_t - \vec{1}m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1}$, where Φ_{1L} and Φ_{2L} are matrix of functions of lagged indicators.

On the other hand, the $\Phi_{1L}(p_t - \vec{1}m_{t-1})$ can readily contain both the aforementioned error-correction mechanisms. The endogenous error-correction mechanism (with various lags, e.g. $p_{1,t-1} - m_{t-1}$, where $p_{1,t-1}$ is the spot price of market 1.) can be included by allowing the diagonal term of Φ_{1L} and Φ_{2L}. For example, when Φ_{1L}=$\begin{pmatrix} a \times L & 0 \\ 0 & b \times L \end{pmatrix}$ and Φ_{2L}=$\begin{pmatrix} a \times L & 0 \\ 0 & b \times L \end{pmatrix}$, then $\Phi_L(p_t, m_t) = \Phi_{1L}(p_t - \vec{1}m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1} = \begin{pmatrix} a(p_{1,t-1} - m_{t-1}) \\ b(p_{2,t-1} - m_{t-1}) \end{pmatrix}$, capturing the endogenous pricing-errors. For a cointegration type of error-correction mechanism (with various lags, e.g. $p_{1,t-1} - p_{2,t-1}$, where $p_{1,t-1}$ is the spot price of market 1 and $p_{2,t-1}$ is the spot price of market 2), can also be readily constructed within this term by calculating the difference between two endogenous error-correction terms. For example, when Φ_{1L}=$\begin{pmatrix} a & -a \\ b & -b \end{pmatrix}$, $\Phi_{1L}(p_t - \vec{1}m_{t-1}) = \begin{pmatrix} a(p_{1,t} - p_{2,t}) \\ b(p_{1,t} - p_{2,t}) \end{pmatrix}$, capturing the cointegration-type pricing-error. Hence, both error-correction type mechanisms are incorporated, and can be distinguished since the non-diagonal term of Φ_{1L} is only related with

[3]The two cointegrated markets' spot prices difference.

cointegration type error-correction terms, with the rest being endogenous error-correction mechanism terms.

2.2 Identification of the General Framework

In this part, we discuss the identification of our framework. The main text provides a summary with full details given in the appendix.
Firstly, we provide the identification constraints of the estimation process. We can get the representation of Δp_t from equation(1) as following:

$$\Delta p_t = \vec{1}Lr_t + (I-L)(I-\Phi_{1L})^{-1}(\vec{1}-\Phi_{2L}\vec{1})r_t + (I-L)(I-\Phi_{1L})^{-1}e_t \quad (2)$$

Where I is a K-by-K identification matrix, L is the lagging indicator and $\Delta p_t = p_t - p_{t-1}$.
Following our model assumptions, the second moment of the asset return provides our information constraint and is represented as follows:

$$\begin{aligned} E[\Delta p_t \Delta p_{t-h}] = E[\sum_{q=0}^{h+1} A_q B_{h+1-q} + \sum_{p=h}^{\infty}(\sum_{q=0}^{p} A_q B_{p-q})(\sum_{q=0}^{p-h} A_q B_{p-h-q})]\vec{1}\sigma_{r_t}^2 + \\ \sum_{p=h}^{\infty} A_p A_{p-h}\Omega \end{aligned} \quad (3)$$

where $\sigma_{r_t}^2$ is the variance of the efficient price shock r_t, Ω is the variance-covariance matrix of the transitory noise e_t, and

$$\begin{aligned} (I-\Phi_{1L})^{-1} &= \phi_{0,1} + \phi_{1,1}L + \phi_{2,1}L^2 + ... \\ \Phi_{2L} &= \psi_{0,2} + \psi_{1,2}L + \psi_{2,2}L^2 + ...\psi_{n,2}L^n \\ A_0 &= \phi_{0,1} \\ A_m &= \phi_{m,1} - \phi_{m-1,1} \\ B_0 &= I - \psi_{0,2} \\ B_m &= \psi_{m,2} \end{aligned} \quad (4)$$

The detailed derivation is in the appendix Section 1.1.

2.2.1 Identification analysis

Due to the complexity of the identification analysis of the most generalized framework, we discuss the identification of a simplified model – the illustrative case with over/under-reaction of efficient price shock and lag 1 of error-correction mechanisms. We believe this simplified derivation is enough because (1) we argue that even if only the lag one error-correction terms are considered, this special case model already captures the further lagged error-correction mechanisms with reasonable structure. The derivation is in the appendix section 1.2, where we also discuss what endogenous error-correction terms capture; and (2) this identification analysis process can also be easily implemented for most of the special cases of our generalized framework.
Again we provide the summary here, the detailed derivation is in the appendix section 1.3.
We consider the model below:

$$p_t = \vec{1} m_t + \alpha r_t + \Phi(p_{t-1} - \vec{1} m_{t-1}) + e_t \tag{5}$$

For the K market case, p_t is a K×1 array of asset prices, $\vec{1}$ is a K×1 vector of 1, m_t is the efficient price scalar, Φ is a K×K matrix (Notice, since we already restricted the lag structure, Φ is a matrix of coefficients not a function of coefficients and lagging indicators), α is a K×1 array of coefficients, and e_t is a K×1 array of noise.
From the model above, we can write out the constraints as follows:

$$\begin{aligned} E(\Delta p_t \Delta p_t') &= [(\vec{1} + \alpha)(\vec{1} + \alpha)' + \alpha\alpha']\sigma_{rt}^2 + 2\Omega \\ E(\Delta p_t \Delta p_{t-h}') &= \Phi^{h-1}(\Phi - I)[(\alpha\vec{1}' + \alpha\alpha'\Gamma(1)')\sigma_{r_t}^2 + \Omega\Gamma(1)'] \end{aligned} \tag{6}$$

In general, we can summarize the identification conclusion as follows :
(1) For the single market case, an additional restriction is needed for identification. This could either be the restriction from Andersen et al. (2022) or the Watson restriction.
(2) For two market case, we follow De Jong and Schotman to assume Ω as diagonal matrix. (3) For multi-market cases, there is no identification issue for most general cases.[4]
Specifically, in the multi-market cases, Φ is always over-identified from the infinite count of autocorrelation constraints.

[4]If identification issue exists, the additional identification restriction can be examined following our appendix.

The detailed derivation and proof are in appendix section 1.3. On the other hand, we aware that the general case (K markets) might be complicated, so we provide the derivation of two special cases in corollary 1 and 2.

2.3 Information share framework

In this section, we calculate the information share (IS) for the generalized model following De Jong and Schotman (2010). Again, the detailed derivation is provided in the appendix section 2.
We start from the most generalized model[5]:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \Phi_L(p_t, m_t) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{7}$$

where m_t is the scalar of the efficient price, r_t is the scalar of the efficient price change, and since we are modeling a K market case, p_t is a K $\times$ 1 vector, α is a K $\times$ 1 matrix, I is a K $\times$ K identity matrix, $\vec{1}$ is a K $\times$ 1 vector of 1, Φ_L is a function of matrix with polynomial lagging indicator and linear coefficient with $Cov(r_t, r_{t-i}) = 0$, $Cov(e_t, e_{t-i}) = 0$, $Cov(r_t, e_t) = 0$ and $Cov(r_t, e_{t-i}) = 0$ for all i $\in \mathbb{N}$.
Following De Jong and Schotman (2010), we define the price innovations at time t as $v_t = p_t - \vec{1}m_{t-1}$.
Therefore, the price innovation can be represented as follows:

$$v_t = \phi_{0,1}(\vec{1} + \alpha - \psi_{0,2}\vec{1})r_t + \Phi_{other,L}L\vec{1}r_t + \sum_{i=0}^{\infty} \phi_{i,1}L^i e_t \tag{8}$$

where the $\Phi_{other,L}L\vec{1}r_t$ term includes the lagged r_t terms, which are orthogonal to r_t.
Then, following De Jong and Schotman (2010), we run the regression $r_t = \gamma_{ols}v_t + \eta_t$, γ_{ols} can be calculated as follows:

$$\begin{aligned} \gamma_{ols} &= \Upsilon^{-1}Cov(v_t, r_t) \\ &= \Upsilon^{-1}\phi_{01}(\vec{1} + \alpha - \psi_{0,2}\vec{1})\sigma^2 \end{aligned} \tag{9}$$

[5]Notice, we do not restrict the lag structure, Φ is a function of coefficients (linear) and lagging indicators (polynomial)

Then the IS can be calculated as:

$$\begin{aligned} R^2 &= 1 - \frac{\sigma_\eta^2}{\sigma^2} \\ &= \frac{\gamma_{ols}' \Upsilon \gamma_{ols}}{\sigma^2} \\ &= \gamma_{ols}'[\phi_{01}(\vec{1} + \alpha - \psi_{0,2}\vec{1})] \end{aligned} \tag{10}$$

Hence, the IS can be represented as follows

$$\begin{aligned} \gamma_{ols} &= \Upsilon^{-1}\phi_{0,1}(\vec{1} + \alpha - \psi_{0,2}\vec{1})\sigma^2 \\ IS &= \gamma_{ols} \circ \phi_{0,1}(\vec{1} + \alpha - \psi_{0,2}\vec{1}). \\ \Upsilon &= E[v_t v_t'] \end{aligned} \tag{11}$$

Equation (11) is the generalized IS estimator from De Jong and Schotman (2010), where they restrict $\phi_{0,1} = I$ and $\psi_{0,2} = 0$[6]. By the generalized representation equation (1), the framework is consistent with various multivariate market microstructure models.

Due to the generalization of our framework, the scope of IS is slightly different in our framework compared to De Jong and Schotman (2010). One of the most significant differences is that De Jong and Schotman (2010)'s IS is restricted to the range of [0,1] while our estimate is unrestricted. In extreme cases the IS can be estimated outside this range. Since economically the IS must lie within the range [0,1], we adopt a quasi-Bayesian approach to ensure our IS measure falls within this range. In practice, where the global minimum produces an IS outside the [0,1] range, we select the best regional minimum. Indeed, in most of these cases, we can find other regional minimums with IS in the [0.1] range with better MLE.

3 Estimated Models

In this section, we discuss and provide our estimated model. We run our empirical model on the two market cases, where s_t is the log price of the SPY market, and f_t is the log price of the SP500-EMINI market.

[6]We aware that they mention that the transitory noise in their model could be non-diagonal. However, in their paper, they adopt much stronger restrictions in their identification analysis, e.g., $Cov(e_t, r_t) = 0$ and the BN decomposition, and orthogonal e_t to restrict their IS to be in the range [0,1].

3.1 Brief Introduction

Our empirical analysis is based on this general framework but with the adoption of different restrictions:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{12}$$

which is equivalent to:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{13}$$

We estimate with two different restrictions :
Model 1:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and following De Jong and Schotman (2010), assuming e_{st} and e_{ft} are orthogonal.
Model 2:

The matrix $\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$ is unrestricted. We still follow De Jong and Schotman by assuming e_{st} and e_{ft} are orthogonal.
Model 1 is the model proposed by De Jong and Schotman (2010), which only allows the over/under-reaction of the efficient price shock. Model 2 is our generalized model, which allows endogenous and cointegration error-correction mechanisms and over/under-reaction of the efficient price.

3.2 Identification of Estimated Models

The identification of model 1 is thoroughly discussed by De Jong and Schotman (2010). The restriction of the diagonal covariance matrix of transitory

term noise guarantees that the model can be fully identified.
The identification of model 2 follows the identification analysis in section 2.2.1 with p_t restricted to the two market case. Specifically, the detail of the identification process is exactly the same as corollary 1 – the $\begin{pmatrix} \alpha_{11}-1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22}-1 \end{pmatrix}$ matrix can be over-identified from the 'ratio' of different lags of second moment restrictions from equation (6), and the remaining parameters (γ-s, $\sigma^2_{r_t}$, $\sigma^2_{e_s t}$ and $\sigma^2_{e_f t}$) can be identified from all constraints. The detailed derivation is in the appendix section 1.3 and corollary 1.

3.3 Information Share of Estimated Models

The IS of both model 1 and model 2 can be calculated from the general model and adding restrictions on the α matrix. The general model is same as before:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{14}$$

which is equivalent to:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11}-1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22}-1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{15}$$

To calculate the IS in equation (11), we need to calculate the variance-covariance matrix of v_t (Υ). It is calculated following the process of the

generalized IS as follows:

$$\begin{aligned}
\begin{pmatrix} s_t \\ f_t \end{pmatrix} &= \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} \\
&\quad + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\
v_t &= \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\
E[v_t, v_t'] &= \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix}' \sigma^2_{r_t} + \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} Cov \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} \\
&\quad \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix}' + Cov \begin{pmatrix} es_t \\ ef_t \end{pmatrix}
\end{aligned} \tag{16}$$

where the price innovation $v_t = \begin{pmatrix} s_t - m_{t-1} \\ f_t - m_{t-1} \end{pmatrix}$ following De Jong and Schotman (2010).

Here the $Cov \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix}$ is estimated as follows by setting the transitory part of p_t as $t_t = \begin{pmatrix} s_t - m_t \\ f_t - m_t \end{pmatrix}$, and Vec is the function mapping a matrix to

its numerator vector representation:

$$\begin{aligned}
\begin{pmatrix} s_t \\ f_t \end{pmatrix} &= \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \\
&\quad \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\
t_t &= \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} t_{t-1} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\
E[t_t, t_t'] &= \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix}' \sigma_{r_t}^2 + \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} E[t_t, t_t'] \\
&\quad \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix}' + Cov \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\
Vec(E[t_t, t_t']) &= Vec(\begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix}' \sigma_{r_t}^2 + Cov \begin{pmatrix} es_t \\ ef_t \end{pmatrix}) + \\
&\quad (\begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \otimes \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix}) Vec(E[t_t, t_t'])
\end{aligned} \tag{17}$$

The $Cov \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix}$ is calculated by reshaping from its vectorized representation.
Then we calculate the IS from the previous conclusion:

$$\begin{aligned}
\gamma_{ols} &= \Upsilon^{-1} \phi_{0,1} (\begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} - \psi_{0,2} \vec{1}) \sigma^2 \\
IS &= \gamma_{ols} \circ \phi_{0,1} (\begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} - \psi_{0,2} \vec{1}). \\
\Upsilon &= E[v_t v_t']
\end{aligned} \tag{18}$$

where $\phi_{0,1}$ is an identification matrix and $\psi_{0,2} = 0$.

3.4 Parameter restrictions

Before we estimate the model and calculate the IS indicator, we still need to discuss the restriction of parameters for model estimation. The first restriction is the IS within [0,1], as we discussed before. In addition, following

Andersen et al. (2022), our restriction needs to uphold the stationarity condition. However, since in the multivariate case for N markets the stationary condition is equivalent to restricting the analytical solution of the eigenvalue to [-1,1] of the matrix rank up to N, and this can be as complex as restricting the root of a polynomial function with the power of N, which might not be feasible when $N > 4$, we do not pursue the general solution of the stationary condition here, rather we only focus on the specific empirical cases previously discussed.
We discuss the stationary condition of the special case model from equation (12):

$$\begin{aligned}\begin{pmatrix} s_t \\ f_t \end{pmatrix} &= \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} \\ &+ \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\ \begin{pmatrix} s_t - m_t \\ f_t - m_t \end{pmatrix} &= \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix}\end{aligned} \tag{19}$$

Therefore, for the permanent-transitory representation, we require the temporary part to be stationary or follow a random-walk[7]. The equation above can be rewritten as a VAR ($t_t = \begin{pmatrix} s_t - m_t \\ f_t - m_t \end{pmatrix}$):

$$\begin{aligned} t_t &= \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} t_{t-1} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \\ t_t &= - \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} t_{t-1} + \begin{pmatrix} (\gamma_s - 1) r_t + es_t \\ (\gamma_f - 1) r_t + ef_t \end{pmatrix} \end{aligned} \tag{20}$$

Then setting $\epsilon_t = \begin{pmatrix} (\gamma_s - 1) r_t + es_t \\ (\gamma_f - 1) r_t + ef_t \end{pmatrix}$, the model is equivalent to a VAR(1) model, and the stationary restriction is equivalent to:

$$\left| \left(I_2 \lambda + \begin{pmatrix} \alpha_{11} - 1 & \alpha_{12} \\ \alpha_{21} & \alpha_{22} - 1 \end{pmatrix} \right) \right| = 0 \tag{21}$$

[7]This is the case when the root is on the unit circle. We allow this case since Andersen et al. (2022) allow α=0 or 2.

Where I_2 is the identity matrix with rank of 2. For all eigenvalue λ that $|\lambda| \leq 1$.
Therefore, it is equivalent to both roots $|\lambda| \leq 1$:

$$(\lambda + \alpha_{11} - 1)(\lambda + \alpha_{22} - 1) - \alpha_{12}\alpha_{21} = 0 \tag{22}$$

Thus, we obtain the restrictions for our special cases:

Under Model 1:

$$\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Hence equation (22) collapses to $\lambda^2 = 0$. Thus, $\lambda_1 = \lambda_2 = 0$ and the model is stationary.

Under Model 2:

The matrix is $\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix}$ is unrestricted. The stationary condition is equal to equation (22). In this case, the restriction of the αs is rather complex. Therefore, we check whether the coefficients indicate the model is stationary following estimation.
Clearly, here we have the stationary condition of $s_t - m_t$ and $f_t - m_t$. However, we are aware, the linear combination of stationary series might not be stationary. The linear combination of stationary series is stationary if and only if these series are jointly weak stationary. In our model, we observe that $Cov(s_t - m_t, f_{t+h} - m_{t+h})$ and $Cov(s_{t+h} - m_{t+h}, f_t - m_t)$ are only related to h and not to t. Therefore, the jointly weak stationarity condition holds in our model and the condition of both $s_t - m_t$ and $f_t - m_t$ being stationary is a sufficient condition for $s_t - f_t$ to be stationary.

4 Empirical Analysis

Having implemented our approach to dealing with the IS representation and identification issues, we implement our multi-market model using data on the SPY and S&P500 Emini (EMINI).

Table 1 presents some basic information about these two assets:

	SPY ETF	S&P500 EMINI
Unit Size	1/10th of Index	\$50 $\times Index$
Trading Venue	NYSE	CME Globex
Ticker Symbol	SPY	ES
24-Hour Trading	No	Yes
Operating Expenses	0.0945%	None

Table 1 SPY and EMINI contract details

4.1 Data

Due to its high liquidity, we use the SPY trading days of the year 2019. For each day, we include all 6.5 common trading hours (9:30 AM to 16:00 PM) of SPY and EMINI, with tick level frequency. All data is obtained from Refinitiv datascope.
There are various popular high-frequency dataset filters, and we follow Kalev and Duong (2008) and Wallace et al. (2019) to filter the dataset using the following steps:

a. Since we are dealing with a period of whole year, we account for the issue of rolling (Emini) contracts following Kalev and Duong (2008), and Wallace et al. (2019) and generate the future contract price as follows:

i. Only the Emini contract closest to the maturity month is used.

ii. The contract is rolled over to the next when it enters its maturity month.

b. By construction, SPY is $\frac{1}{10}$ of the scale of the S&P500 Index and S&P future contracts. Therefore, following Hasbrouck(2003) and Budish et al. (2015), we scale the SPY by ten.

c. The difference between the scaled SPY and Emini still includes two major discrepancies: the cost-of-carry and the cash component of the ETF. Following Hasbrouck (2003), we allow the constant component in the state-space representation and VAR framework per estimation window. Specifi-

cally, Hasbrouck (2003) argues that it is robust to assume that the difference is constant at the daily frequency. Since we are using the shorter window, this assumption should be reasonable.
Tables 2–3 report the summary statistics of the SPY and the EMINI. We observe that both trading volume and the Amihud Illiquidity measure indicate the SPY market is far more liquid than the EMINI market.

Month	Price	Range	Volume	Amihud Illiquidity (e-8)
Jan, 2019	259.70 (5.90)	3.22 (1.28)	825.20 (199.12)	1.81
Feb, 2019	275.05 (3.80)	1.91 (0.58)	615.78 (103.53)	1.27
Mar, 2019	279.96 (2.49)	2.60 (1.14)	684.86 (143.80)	1.73
Apr, 2019	289.57 (2.53)	1.59 (0.52)	482.27 (85.63)	1.46
May, 2019	285.46 (4.76)	3.14 (1.25)	708.02 (213.76)	1.96
Jun, 2019	288.96 (5.32)	2.22 (0.89)	561.84 (162.95)	1.56
Jul, 2019	298.86 (1.78)	2.00 (1.05)	434.70 (138.03)	2.59
Aug, 2019	289.78 (3.41)	4.13 (2.01)	769.76 (309.42)	2.81
Sep, 2019	298.03 (2.82)	2.35 (1.02)	552.78 (138.94)	1.87
Oct, 2019	297.21 (4.73)	2.52 (1.31)	522.13 (210.54)	2.35
Nov, 2019	310.13 (2.64)	1.49 (0.40)	438.36 (97.63)	1.56
Dec, 2019	317.44 (4.30)	1.66 (0.89)	499.86 (226.49)	1.82

Table 2 Descriptive Statistics of SPY

Note, the table reports the descriptive statistics of SPY. Volume is the average daily traded volume in units of 100,000 and Amihud Illiquidity is the daily average scaled by 10^{-8}. Range is the average daily price range where the range is the daily high price minus the daily low price, and price is the average daily transacted price. The standard deviation of each statistic is in parentheses.

Month	Price/10	Range/10	Volume	Amihud Illiquidity (e-5)
Jan, 2019	260.32 (5.80)	3.26 (1.29)	12.40 (2.61)	1.19
Feb, 2019	275.23 (3.69)	1.90 (0.60)	10.22 (1.88)	0.79
Mar, 2019	280.83 (2.84)	2.63 (1.15)	9.87 (5.34)	4.87
Apr, 2019	290.51 (2.38)	1.61 (0.52)	8.70 (1.54)	0.79
May, 2019	285.68 (4.98)	3.16 (1.26)	14.16 (4.25)	1.01
Jun, 2019	289.43 (5.60)	2.24 (0.89)	6.34 (5.69)	2.06
Jul, 2019	299.80 (1.64)	2.03 (1.06)	8.47 (2.20)	1.39
Aug, 2019	289.94 (3.44)	4.11 (2.03)	14.25 (4.63)	1.48
Sep, 2019	298.33 (2.77)	2.41 (1.05)	7.11 (5.62)	12.35
Oct, 2019	297.77 (4.62)	2.55 (1.33)	9.89 (2.98)	1.21
Nov, 2019	310.23 (2.52)	1.51 (0.41)	8.24 (1.65)	0.82
Dec, 2019	317.88 (4.78)	1.67 (0.90)	5.53 (5.04)	11.6

Table 3 Descriptive Statistics of EMINI

Note, the table reports the descriptive statistics of EMINI. Volume is the average daily traded volume in units of 100,000m, price and range are the average daily traded price in units of 10 and Amihud Illiquidity is daily average scaled by 10^{-5}. Range is the average daily price range where the range is the daily high price minus the daily low price, and price is the average daily transacted price. The standard deviation of each statistic is in parentheses.

4.2 Abnormal Autocorrelation

Following Andersen et al. (2022), firstly, we investigate the autocorrelation pattern or return dynamic in SPY and EMINI. The abnormal autocorrelation structure has been mentioned in various papers (e.g., Hansen and Lunde, 2006). Andersen et al. (2022) indicate by plotting the volatility signature plot that several stocks exhibit abnormal positive autocorrelation on various days. We find a similar pattern in our case. For example, figure one presents two kinds of observed patterns in the volatility signature plots of the SPY[8].

[8]We find similar patterns within the EMINI market.

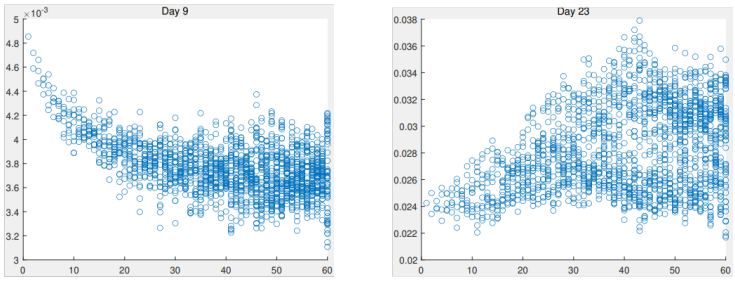

Figure 1: Volatility Signature Plots

The left figure is the volatility plot of the 9th trading day of Sep, and the right figure is the 23rd trading day of Sep. As discussed by Hansen and Lunde (2006) and Andersen et al. (2022), if the autocorrelation of return is always negative, the volatility signature plot should have a downward trajectory (as shown on the left). However, on the right-hand side we observe some evidence of an upward trajectory. This suggests evidence of autocorrelation dynamics within the series. Moreover, we identify some complex volatility signature plot structures. For example, figure 2 plots the 17th trading day of Sep. Here, the volatility signature plot indicates an autocorrelation pattern that is far more complex than the 'all negative' pattern that many previous market microstructure models assume.

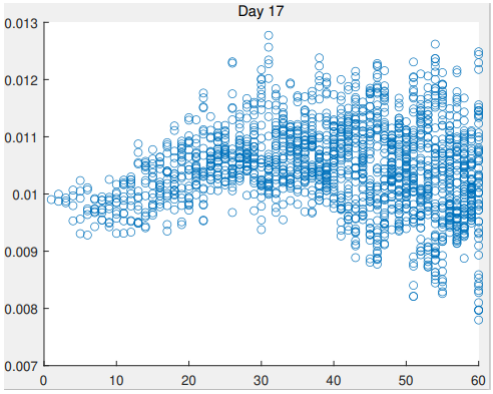

Figure 2: Volatility Signature Plot

This primary evidence suggests that the autocorrelation pattern in the SPY and EMINI markets can be quite complex. Thus, it is important

that any model is able to capture more complex and longer autocorrelation structures. One straightforward way is to incorporate the endogenous error-correction mechanism as indicated by Andersen et al. (2022).

4.3 Statistical Results

The log-likelihood-ratio test is used to examine whether our unrestricted model 2 (capturing the error-correction mechanisms and transitory noise correlation) fits the data better than restricted model 1 (De Jong and Schtoman (2010) model). We find that out of 9777 windows, the log-likelihood ratio test indicates that in 9746 windows, model 2 are preferred compared to model 1[9], which indicates that more than 99% of the time, model 2 is statistically better than the De Jong and Schotman (2010) specification, model 1. Table 4 reports the intraday pattern of the likelihood ratio test results. It is observed that model 2 consistently outperforms model 1.

	Percentage that model 2 is preferred
Average	99.17%
Min	98.02%
Pct50	98.81%
Max	100.00%

Table 4 Descriptive Statistics of Likelihood Ratio Test

Note, the table reports the descriptive intraday statistics of the Likelihood Ratio test. The descriptive statistics were based on the estimation/LR test result regrouped by the period. For example, the first group contains the estimation/LR test result of 9:30 AM-9:40 AM on all days.

4.4 Model Estimation Result

For estimation, we split the tick level data of full 6.5 common trading hours (9:30 AM to 16:00 PM EST) of SPY (s_t) and EMINI (f_t) by every 10 minute , giving a total of 39 trading windows each day. The intraday pattern plots

[9]We remove the window Jul 3, Nov 29 and Dec 24 after 12:15 due to missing data.

are calculated by average for each 10 minute intraday window across all days in the sample. For example, for the intraday pattern of 9:30 AM to 9:40 AM, we calculate the average of the estimates from all 252 trading days for the window 9:30 AM to 9:40 AM.
Recalling equation (12), we can separate out the two error-correction mechanisms and rewrite the equation as follows:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} + \alpha_{12} & 0 \\ 0 & \alpha_{21} + \alpha_{22} \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{12} \times (s_{t-1} - f_{t-1}) \\ \alpha_{21} \times (f_{t-1} - s_{t-1}) \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{23}$$

where s_t is the log spot price of the spot market SPY at time t, f_t is the log spot price of the EMINI futures market at time t, and m_t is the efficient price at time t.
From equation (23), γ_s and γ_f capture the spot price reaction to the efficient price change within the SPY and EMINI markets. When γ_s or γ_f is in in the range (0,1), the underlying market partially reflects the efficient price change. When γ_s or γ_f is 1, the underlying market perfectly reflects the efficient price change, and if either is 0, the efficient price change is not reflected at all immediately. Therefore for most cases we expect γ_s and γ_f be estimated in the range [0,1]. $(\alpha_{11} + \alpha_{12})$ and $(\alpha_{21} + \alpha_{22})$ capture the endogenous error-correction mechanism. When $(\alpha_{11} + \alpha_{12})$ or $(\alpha_{21} + \alpha_{22})$ is in the range (0,1), it means the underlying market is partially correcting its own historical pricing error. When $(\alpha_{11} + \alpha_{12})$ or $(\alpha_{21} + \alpha_{22})$ is 0, it means the underlying market does not correct its own historical pricing-error at all, and either term is 1, its own historical pricing-error is fully corrected. Equivalently, therefore, we expect most estimates of $(\alpha_{11} + \alpha_{12})$ and $(\alpha_{21} + \alpha_{22})$ to be in the range [0,1]. Lastly, α_{12} and α_{21} captures the cointegration type error-correction mechanism. When α_{12} or α_{21} in the range (-1,0), it means the underlying market past cointegration pricing-error is partially corrected. Similarly, when α_{12} or α_{21} is 0, it means the price difference between these markets (cointegration pricing-error) is not corrected at all, and either term is -1, the price difference is fully corrected. Therefore, we expect most α_{12} and α_{21} in the range [-1,0].

4.4.1 Estimation of Multivariate Framework without Error-Correction Mechanisms

In this section, we provide the estimation of model 1 (De Jong and Schotman (2010)'s framework). Table 5 reports the summary statistics fo the estimation. Generally, the mean estimation of each parameter is in line with expectations, though estimation of higher moments indicates some potential instability in parameter distributions.

	γ_s	γ_f	Std(r_t)	std(es_t)	std(ef_t)
Mean	0.76	0.80	3.25E-05	1.13E-05	2.32E-05
Std	0.24	0.24	1.95E-05	7.19E-06	1.48E-05
Skewness	-0.91	-1.07	1.92	24.94	56.10
Kurtosis	3.09	3.63	9.17	953.32	3911.48
Pct1	0.08	0.09	8.68E-06	1.70E-06	1.55E-05
Pct10	0.44	0.50	1.40E-05	7.17E-06	1.90E-05
Pct25	0.53	0.60	1.90E-05	8.99E-06	2.06E-05
Pct50	0.85	0.90	2.76E-05	1.10E-05	2.25E-05
Pct75	0.94	0.98	4.04E-05	1.30E-05	2.47E-05
Pct90	1.00	1.03	5.72E-05	1.50E-05	2.70E-05
Pct99	1.09	1.14	1.02E-04	2.23E-05	3.43E-05

Table 5 Estimation of model 1

In figure 3 we report each market's reaction to an efficient price change γ from De Jong and Schotman (2010)'s model. The parameter distributions are not very stable, but it can still be observed that the scale of γ_s and γ_f are roughly the same. This suggests that, as expected, the SPY and EMINI market react to the efficient price shock in similar ways.

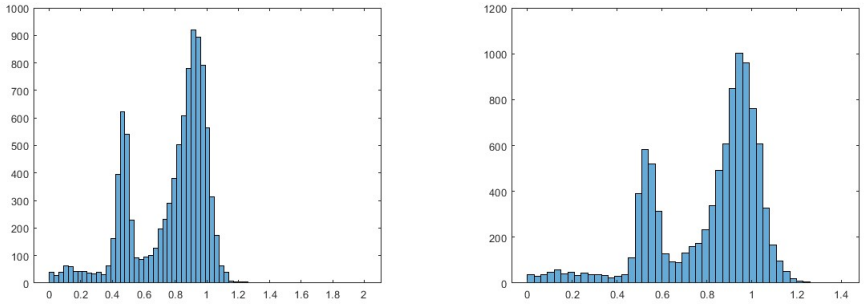

Figure 3: γ_s and γ_f of model 0

Similarly, the scale and the distribution of the volatility estimation from De Jong and Schotman (2010)'s model are also in line with expectations. Figure 4 presents the standard deviation of efficient price change and the transitory noise terms of two markets. Both the scales and the distributions are as expected.

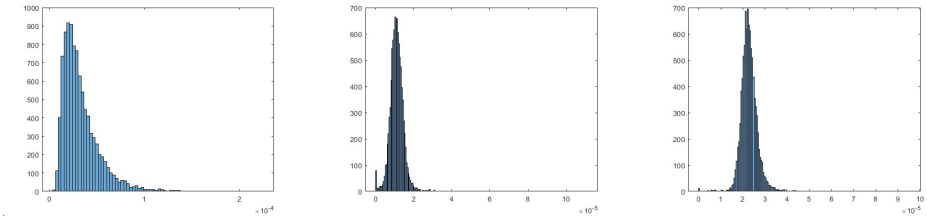

Figure 4: Std(r_t), Std(es_t) and Std(ef_t) of model 1

The distribution pattern of the information share of the SPY and EMINI markets are presented in figure 5. The plots suggest the IS of the two markets does not match prior expectations and previous literature. Normally, the IS (or the efficiency of the reaction to an efficient price change) should be similar for these two markets since both markets have their relative advantages: The SPY market attracts more investors due to its very high liquidity (from table 2 and table 3), while the EMINI market is naturally attractive to informed investors due to leverage. Therefore, we expect the IS of these two markets should be similar. However, figure 5 indicates that the SPY market reacts to the efficient price change more efficiently comparing to the EMINI market. We believe this is inconsistent with the common understanding of these markets.

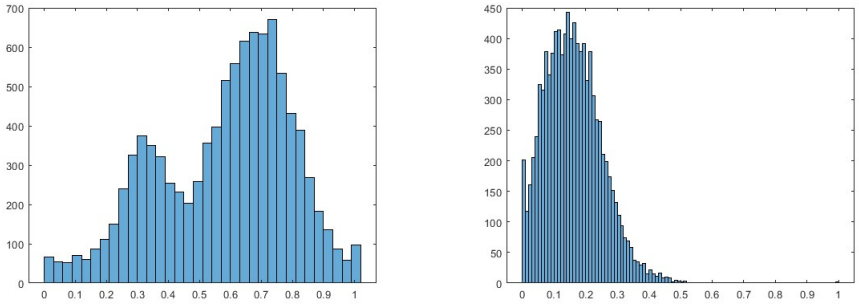

Figure 5: SPY and EMINI IS

Moreover, roughly speaking, the IS is mainly determined by the coefficient of γ_s, γ_f, and the inverse of the volatility of the transitory noise ((es_t) and (ef_t)). In other words, for each market, the closer γ_s or γ_f are to 1, and the smaller the transitory noise terms are, the larger the IS. However, as indicated, the scale of γ_s and γ_f are similar, as are the two transitory noises, yet the ISs of these two markets are very different. We believe this indicates that model 1 does not perform well in this case due to the strong assumption that the two markets are only correlated through the efficient price change. This assumption does not ideally fit the case selected.

4.4.2 Estimation of Multivariate Framework with Error-Correction Mechanisms

In this section, we provide the estimation of our model that accounts for the over/under-reaction of the efficient price change, endogenous and cointegration type error-correction mechanisms, the volatility of the efficient price change and transitory noise of the two markets, and the co-movement captured, but unexplained, by our framework.

Table 6 reports the summary statistics of our model 2 estimation. The majority of the statistics fit with expectations. Moreover, the higher moment statistics indicate the distributions of model 2 estimation are more stable than the distributions of model 1 estimation.

	γ_s	γ_f	$\alpha_{11}+\alpha_{12}$	α_{12}	α_{21}	$\alpha_{21}+\alpha_{22}$	Std(r_t)	std(es_t)	std(ef_t)
Mean	0.94	0.99	0.13	-0.11	-0.06	0.65	3.31E-05	8.85E-06	1.62E-05
Std	0.10	0.13	0.30	0.14	0.43	0.64	2.07E-05	4.94E-06	6.24E-06
Skewness	-0.97	-0.42	0.44	-0.64	0.45	0.23	6.00	5.08	3.86
Kurtosis	12.31	13.18	3.76	6.27	2.97	2.97	164.31	101.30	89.00
Pct1	0.64	0.61	-0.56	-0.54	-0.93	-0.63	8.77E-06	4.56E-07	2.17E-06
Pct10	0.81	0.88	-0.21	-0.29	-0.58	-0.23	1.45E-05	3.42E-06	8.95E-06
Pct25	0.89	0.94	-0.03	-0.19	-0.36	0.26	1.96E-05	5.98E-06	1.27E-05
Pct50	0.96	0.99	0.07	-0.10	-0.04	0.66	2.81E-05	8.60E-06	1.65E-05
Pct75	1.00	1.04	0.29	-0.02	0.12	0.97	4.09E-05	1.14E-05	1.97E-05
Pct90	1.04	1.13	0.55	0.03	0.60	1.54	5.79E-05	1.42E-05	2.25E-05
Pct99	1.16	1.35	0.94	0.19	0.97	2.20	1.03E-04	2.02E-05	2.91E-05

Table 6 Estimation of model 2

In terms of the reaction of each market to a change in the efficient price and how γ evolves across time, figure 6 reports the distributions of γ_s and γ_f while figure 7 plots the intraday patterns of γ_s and γ_f. As shown in equation (23), $\gamma = 0$ indicates that the efficient price change is not reflected within the underlying asset price while $\gamma = 1$ indicates the efficient price change is fully reflected within the underlying asset price. Therefore, for the SPY and EMINI markets, we observe the efficient price change is mainly reflected within both price series (under-reaction to the efficient price change occurs when $\gamma < 1$ and over-reaction when $\gamma > 1$). This echoes our expectation because, in the year 2019, both these markets are very efficient regarding the information shock.

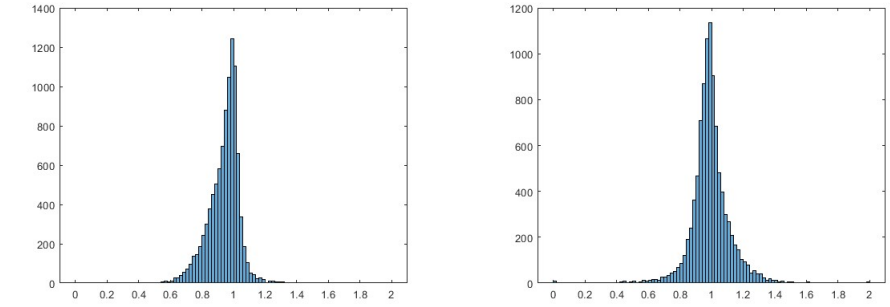

Figure 6: γ_s and γ_f

We observe that the distributions of γ_s and γ_f from model 2 more closely represent a normal distribution compared with the estimations of model 1. The intraday pattern of γ_s indicates that shortly after the SPY market opens, the asset price is more sensitive to an efficient shock and related investors are more active. It suggests that the SPY market is more information-driven during this period. On the other hand, the EMINI market is typically very sensitive to efficient price change since its γ_f is always around 1. Shortly after SPY market opens and before it closes, the EMINI market is more efficiently reflecting the efficient price change comparing to other periods. Overall, generally, both markets are pretty efficiently reflecting the efficient price.

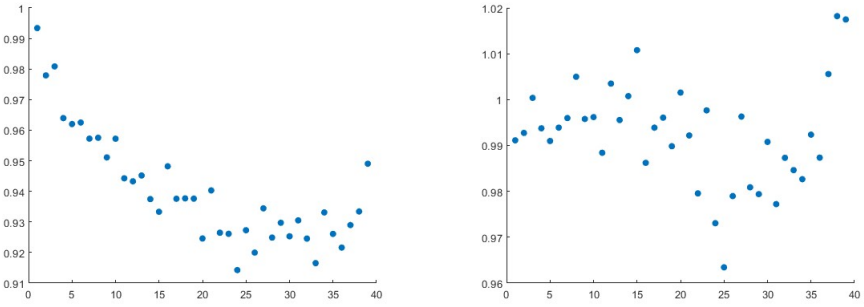

Figure 7: Intraday γ_s and γ_f

As shown in equation (23), $\alpha_{11} + \alpha_{12}$ captures the endogenous error-correction mechanism within the SPY market, and $\alpha_{21} + \alpha_{22}$ captures the

endogenous error-correction mechanism within the EMINI market. When either of these terms is 0, it means the past pricing error regarding its own historical price is not corrected. When the estimated coefficients sum to 1, then its own past pricing error is fully corrected.
Figure 8 reports the distributions of the endogenous error-correction mechanisms as captured by $\alpha_{11} + \alpha_{12}$ and $\alpha_{21} + \alpha_{22}$. Figure 9 reports the intraday patterns of $\alpha_{11} + \alpha_{12}$ and $\alpha_{21} + \alpha_{22}$.

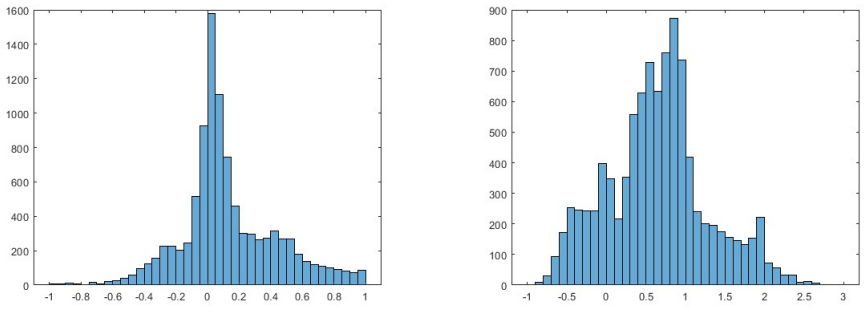

Figure 8: $\alpha_{11} + \alpha_{12}$ and $\alpha_{21} + \alpha_{22}$

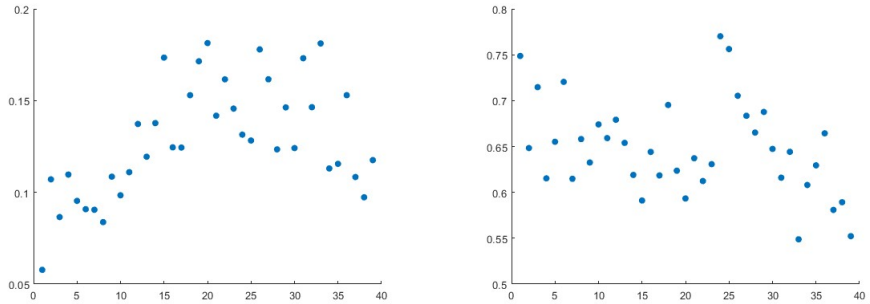

Figure 9: Intraday $\alpha_{11} + \alpha_{12}$ and $\alpha_{21} + \alpha_{22}$

It is observed that figure 8 and figure 9 indicate that most of the time, both the SPY and EMINI are partially correcting their past individual pricing error. Though, the proportion of past pricing-error being corrected, is larger in EMINI market than in the SPY market. This is consistent with expectations, since the EMINI market is more leveraged, and investors are therefore

more actively hunting the pricing error within its own historical price. Our findings suggest endogenous pricing-error seekers are continuously active in both the SPY and EMINI markets during the entire trading period (every 10 minute window), with investors in the EMINI market more efficiently correcting past pricing errors compared to SPY investors. On intraday level, our estimation indicates that the SPY market is more efficiently correcting its past pricing-error during middle of the day, and the EMINI market is more actively correcting its past pricing error shortly after the SPY market opens and in the middle of the day.

Again, as shown in equation (23), the second type of error-correction mechanism in our model – the cointegration type error-correction mechanism, is captured by α_{12} in the SPY market, and by α_{21} for the EMINI market. When either α_{12} or α_{21} is 0, the price difference between the SPY and EMINI is rarely corrected with the respective underlying market. When either coefficient is -1, the two-market pricing error is fully corrected within the respective underlying market. Similar to earlier results, we present the distributions of α_{12} and α_{21} in figure 10 and the intraday patterns in figure 11.

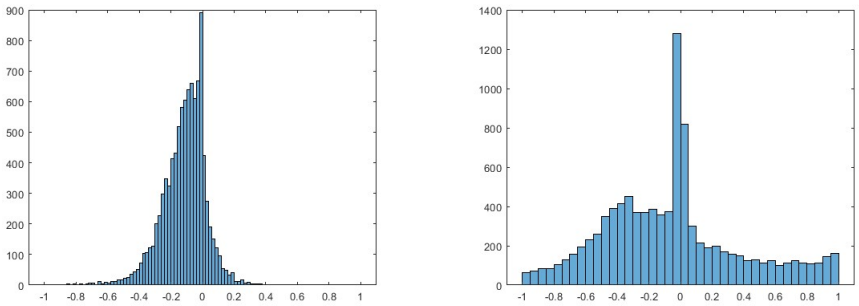

Figure 10: α_{12} and α_{21}

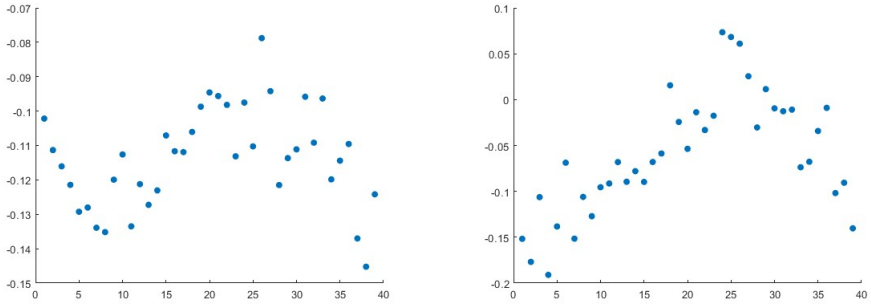

Figure 11: Intraday α_{12} and α_{21}

Figure 10 (left panel α_{12} – SPY, right panel α_{21} – EMINI) indicates that the cointegration error-correction mechanism plays an important role in both the SPY and EMINI markets. This suggests that the SPY and EMINI markets are learning from each other.
The intraday pattern in Figure 11 indicates the existence of a 'highly information-driven period' (shortly after the opening and before the closing of the SPY market) when investors are more actively searching the price difference between these two markets, and the cointegration type error-correction mechanism is more active. This is represented by the two inverse 'U' shapes in the figure.[10]

Figure 12 reports the distribution and figure 13 plots the intraday pattern of the standard deviation of efficient price change r_t, and the transitory noise terms of the SPY and EMINI markets (e_{st} and e_{ft}).

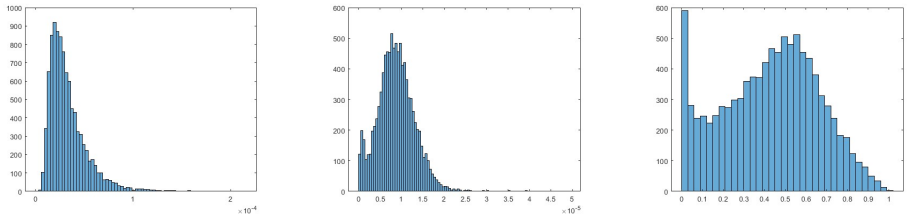

Figure 12: Std(r_t), Std(es_t) and Std(ef_t)

[10]We aware that from 9:30-10:00 AM the cointegration error-correction mechanism is less active than 10:00-10:30AM, but still the first 30 minutes cointegration error-correction mechanism is pretty active comparing the average of the day.

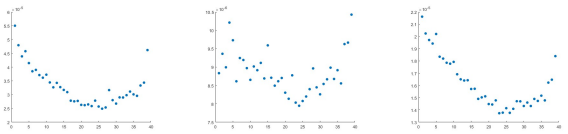

Figure 13: Intraday Std(r_t), Std(es_t) and Std(ef_t)

From figure 13 it can be observed that both these markets receive more information shortly after the opening and before the closing of the SPY market, reflected by higher efficient price shock volatility. Moreover, as expected, the transitory noise in both markets is stronger during these periods due to the presence of more active investors and more volatile markets.

The intraday pattern of the information share in the SPY (left panel) and EMINI (right panel) markets is presented in figure 14. The IS of each of the two markets are similar, meaning the SPY and EMINI markets are roughly equally actively reflecting the efficient price change. This echos our earlier result, especially with respect to the estimation of γ_s and γ_f. Moreover, the intraday pattern of the SPY and EMINI ISs is also consistent with our earlier findings. At the opening of the SPY market, the information share is significantly higher than at other periods, so is γ_s. Conversely, the information share from the EMINI market, γ_f, is higher shortly before the opening and closing of the SPY market. This pattern indicates that shortly after the opening of the SPY market, due to the inflow of information, both SPY and EMINI reflect the efficient price change more actively compared to other periods or is more sensitive to information than during the rest of the day due to the more active investors. During the closing of the SPY market, the EMINI market becomes more sensitive to efficient price changes or information shocks. In general, the IS estimation echoes our previous estimation – the two markets have similar ISs, but each market is active at a different point in the trading day.[11].

[11]The IS is more closely related to the γ_s and γ_f (over/under-reaction to the efficient price change), rather than the two error-correction mechanisms

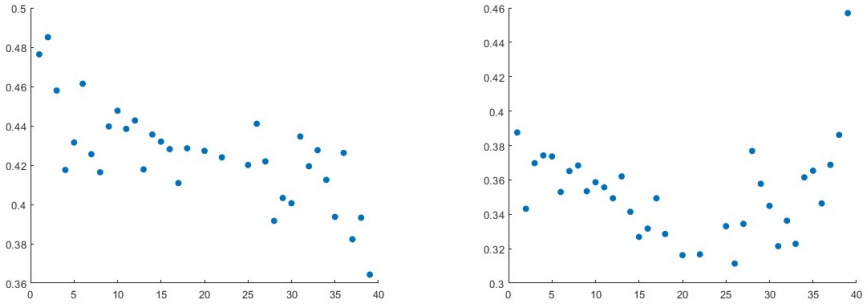

Figure 14: SPY and EMINI IS

5 Conclusion

We propose a framework as a generalization of both the endogenous pricing-error model of Andersen et al. (2022) and the IS framework of De Jong and Schotman (2010) to establish a richer model at the microstructure level to better measure information shares. We make multiple methodological and empirical contributions to the literature. (1) Our generalized framework provides a solution to the identification issues associated with the model of Andersen et al. (2022). (2) We generalized the IS framework of De Jong and Schotman (2010). (3) Our framework fills the gap between various market microstructure models and IS estimation by extending the IS from Hasbrouck (1995) and De Jong and Schotman (2010). (4) The intraday pattern of various market mechanisms (over/under-reaction of the efficient price change, and endogenous and cointegration error-correction mechanisms) and the IS can be captured with our framework more accurately without higher frequency dataset and computational power. (5) We provide an example that when the endogenous pricing-error terms are included in the model, the model can capture further second moment/autocorrelation constraints. Our empirical analysis highlights that (1) Both the SPY and EMINI markets actively reflect the efficient price change. At the intraday level, both of them are more efficient shortly after the opening of the SPY market, and the EMINI market is also very efficient shortly before the closing of SPY market. (2) Both the SPY and EMINI markets are partially correcting their past endogenous pricing-error. Shortly after the opening and before the closing of the SPY market, both SPY and EMINI have more active endogenous

error-correction mechanisms. Compared to the SPY market, EMINI is more efficient in correcting its own historical/endogenous pricing-error. (3) The cointegration pricing-error are partially corrected in both the SPY and EMINI markets. At the intraday level, for both SPY and EMINI markets, the cointegration error-correction mechanism is more active shortly after the SPY market opens and shortly before the close of the SPY market. (4) During our estimation period, SPY and EMINI's ISs are found to be similar, indicating that these two markets are roughly equally active in reflecting the efficient price dynamics or information shock. At the intraday level, SPY is more efficient shortly after it opens comparing other times, and the EMINI market is more efficient shortly after the opening and before the closing of SPY market. These patterns echo the intraday pattern of the over/under-reaction of the efficient price change coefficient. (5) As expected, there is more information flow shortly after the opening and before the close of the SPY market, which is captured by the volatility of the efficient price change. The transitory noise is also larger for both the SPY and EMINI markets in these two time periods. Overall, our proposed framework highlights the importance of capturing the different dynamics and mechanisms at play in the multiple market setting. We observe notable intraday patterns that further underline the importance of a better and richer understanding of intraday dynamics, mispricing, and learning mechanisms and how these impact price discovery and information shares.

6 Reference

Andersen, T., Bollerslev, T., Diebold, F., Labys, P., 2000. Great realisations. Risk Mag. 18, 105–108.

Andersen, T. G., Archakov, I., Cebiroglu, G., Hautsch, N. (2022). Local mispricing and microstructural noise: A parametric perspective. Journal of Econometrics.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., Shephard, N. (2008). Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica, 76(6), 1481-1536.

Booth, G. G., So, R. W., Tse, Y. (1999). Price discovery in the Ger-

man equity index derivatives markets. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 19(6), 619-643.

Budish, E., Cramton, P., Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. The Quarterly Journal of Economics, 130(4), 1547-1621.

Cabrera, J., Wang, T., Yang, J. (2009). Do futures lead price discovery in electronic foreign exchange markets?. Journal of Futures Markets: Futures, Options, and Other Derivative Products, 29(2), 137-156.

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

De Jong, F., Schotman, P. C. (2010). Price discovery in fragmented markets. Journal of Financial Econometrics, 8(1), 1-28.

Duong, H. N., Kalev, P. S. (2008). The Samuelson hypothesis in futures markets: An analysis using intraday data. Journal of Banking Finance, 32(4), 489-500.

Engle, R. F., Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. Econometrica: journal of the Econometric Society, 251-276.

Forte, S., Pena, J. I. (2009). Credit spreads: An empirical analysis on the informational content of stocks, bonds, and CDS. Journal of Banking Finance, 33(11), 2013-2025.

Hansen, P. R., Lunde, A. (2006). Realized variance and market microstructure noise. Journal of Business Economic Statistics, 24(2), 127-161.

Harris, F. H. D., McInish, T. H., Shoesmith, G. L., Wood, R. A. (1995). Cointegration, error correction, and price discovery on informationally linked security markets. Journal of financial and quantitative analysis, 30(4), 563-579.

Harris, F. H. D., McInish, T. H., Wood, R. A. (2002). Security price ad-

justment across exchanges: an investigation of common factor components for Dow stocks. Journal of financial markets, 5(3), 277-308.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. The journal of Finance, 50(4), 1175-1199.

Hasbrouck, J. (2003). Intraday price formation in US equity index markets. The Journal of Finance, 58(6), 2375-2400.

Hasbrouck, J. (2021). Price discovery in high resolution. Journal of Financial Econometrics, 19(3), 395-430.

O'hara, M. (2015). High frequency market microstructure. Journal of financial economics, 116(2), 257-270.

Lütkepohl, H. (2005). New introduction to multiple time series analysis. Springer Science Business Media.

Putniņš, T. J. (2013). What do price discovery metrics really measure?. Journal of Empirical Finance, 23, 68-83.

Wallace, D., Kalev, P. S., Lian, G. (2019). The evolution of price discovery in us equity and derivatives markets. Journal of Futures Markets, 39(9), 1122-1136.

Vives, X. (2010). Information and learning in markets: the impact of market microstructure. Princeton University Press.

Yan, B., Zivot, E. (2010). A structural analysis of price discovery measures. Journal of Financial Markets, 13(1), 1-19.

Mispricing, Learning, and Price Discovery - Appendix

1 Identification of the General Framework

In this part, we discuss the identification of the generalized model below.

$$\begin{aligned} p_t &= \vec{1} m_t + \alpha r_t + \Phi_L(p_t, m_t) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{1}$$

where m_t is the scalar of efficient price at time t, r_t is the scalar of efficient price change at time t. For the general K cointegrated markets case, p_t is a K $\times$ 1 vector contains the spot price of K markets, α is a K $\times$ 1 vector of coefficients, $\vec{1}$ is a K $\times$ 1 vector of 1, Φ_L is a matrix of linear combination of lagged p_t and m_t. e_t is a K $\times$ 1 vector of normal distributed noise terms. In our framework, we restrict $Cov(r_t, r_{t-i}) = 0$, $Cov(e_t, e_{t-i}) = \{0\}$, $Cov(r_i, e_j) = \vec{0}$ for all i, j $\in \mathbb{N}$[1], where $\{0\}$ is a K $\times$ K matrix of 0 and $\vec{0}$ is a K $\times$ 1 vector of 0.

1.1 Identification Restrictions

Firstly, we discuss the identification constraints of estimation. Easily, eq.(1) can be written as following by decomposing $\Phi_L(p_t, m_t) = \Phi_{1L}(p_t - \vec{1} m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1}$. Therefore, the eq(1) can be written as following:

$$\begin{aligned} p_t &= \vec{1} m_t + \Phi_{1L}(p_t - \vec{1} m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1} + e_t \\ (I - \Phi_{1L})(p_t - m_{t-1}) &= (\vec{1} - \Phi_{2L}\vec{1}) r_t + e_t \\ p_t - m_{t-1} &= (I - \Phi_{1L})^{-1}(\vec{1} - \Phi_{2L}\vec{1}) r_t + (I - \Phi_{1L})^{-1} e_t \\ \Delta p_t &= \vec{1} L r_t + (I - L)(I - \Phi_{1L})^{-1}(\vec{1} - \Phi_{2L}\vec{1}) r_t \\ &\quad + (I - L)(I - \Phi_{1L})^{-1} e_t \end{aligned} \tag{2}$$

Where I is K-by-K identity matrix, L is the lagging indicator and $\Delta p_t = p_t - p_{t-1}$.
To analyze the return dynamic of the model above, we need to decompose the $(I - \Phi_{1L})^{-1}$ and Φ_{2L}.
Since Φ_{2L} contains lagging indicator, we can write Φ_{2L} as $\Phi_{2L} = \psi_{0,2} + \psi_{1,2} L +$

[1] Here 0 is not included.

$\psi_{2,2}L^2 + ...\psi_{n,2}L^n$ by reorganizing based on lagging terms[2].
For $(I-\Phi_{1L})^{-1}$, the decomposition would be slightly more complex and include potentially infinite count of terms. In this case, we assume $||\Phi_{1L}|| < 1$. The infinite term representation is normally needed because of the inverse of lagging metrics. It can be achieved by unfolding the following decomposition to factorization from the matrix of polynomial of L to the product of the highest power of 1. Therefore, we can move the lagging indicators out, and Ψ_{i1} matrices do not include lagging indicator L. [3]
Specifically, here we firstly decompose $(I-\Phi_{1L})^{-1}$ to finite amount of "factors" and then decompose each of these finite terms into infinite terms as follows, where Φ_{i1} does not contain lagging term:

$$\begin{aligned}(I-\Phi_{1L})^{-1} &= \Phi_{0,1}(I-\Phi_{11}L)^{-1}(I-\Phi_{21}L)^{-1}...(I-\Phi_{i1}L)^{-1} \\ &= \Phi_{0,1}(I+\Phi_{11}L+\Phi_{11}^2L^2+...)(I+\Phi_{21}L+\Phi_{21}^2L^2+...)... \\ &(I+\Phi_{i1}L+\Phi_{i1}^2L^2+...) \\ &= \Phi_{0,1}\prod_{k=1}^{i}\sum_{j=1}^{\infty}(I+\Phi_{k1}^jL^j)\end{aligned} \tag{3}$$

On the other hand, we rewrite the eq.(3) as following:

$$\begin{aligned}(I-\Phi_{1L})^{-1} = \phi_{0,1} + L\sum_{p_1\in[1,i]}\Phi_{p_11} + L^2\sum_{p_1,p_2\in[1,i]}\Phi_{p_11}\Phi_{p_21} + \\ L^3\sum_{p_1,p_2,p_3\in[1,i]}\Phi_{p_11}\Phi_{p_21}\Phi_{p_31} + ...\end{aligned} \tag{4}$$

Then this equation can be simplified as:

$$\begin{aligned}(I-\Phi_{1L})^{-1} &= \phi_{0,1} + \phi_{1,1}L + \phi_{2,1}L^2 + ... \\ \phi_{m,1} &= \sum_{p_q\in[1,i]}\prod_{q=1}^{m}\Phi_{p_q1}\end{aligned} \tag{5}$$

where m,q as integer and p_m not necessarily different from p_n for $m \neq n$.
To distinguish these two kinds of decomposition, for infinite decomposition (for $(I-\Phi_{1L})^-1$) we use ϕ as coefficient metrics, and for finite decomposition (for $(I-\Phi_{2L})$) we use ψ for the decomposed coefficient metrics (there is no lagging term embedded in these matrices).
Since we assume that the noise is normal distributed, our constraints would be the second moment conditions.
Before we calculate the moment restrictions of E[p_t, p_{t-h}'], we rewrite out the

[2]Finite count of terms

[3]For example, for a one-by-one matrix of $(I-\Phi_{1L})^{-1} = [(1-3L+2L^2)^{-1}]$ is decomposed to $(1-2L)^{-1}(1-L)^{-1}$. Then $(I-\Phi_{11}L) = 1-2L$ and $(I-\Phi_{21}L) = 1-L$. In this way, the Ψ_{i1} matrices do not include the lagging indicator L.

Δp_t with following setting to simplify the equation above:

$$\begin{aligned} A_0 &= \phi_{0,1} \\ A_m &= \phi_{m,1} - \phi_{m-1,1} \\ B_0 &= I - \psi_{0,2} \\ B_m &= -\psi_{m,2} \end{aligned} \tag{6}$$

Then the Δp_t as following and splitting the p_t to two parts, the lagging terms of $r_t - m_t - m_{t-1}$ and e_t of time from (1) t-h+1 to t and (2) the others with higher lags from eq.(2):

$$\begin{aligned} \Delta p_t &= \vec{1} L r_t + (A_0 + A_1 L + A_2 L^2 + ...)(B_0 + B_1 L + B_2 L^2 + ...)\vec{1} r_t + \\ &\quad (A_0 + A_1 L + A_2 L^2 + ...) e_t \\ &= \vec{1} L r_t + \sum_{p=0}^{\infty} \sum_{q=0}^{p} A_q B_{p-q} + \sum_{r=0}^{\infty} A_r L^r e_t \\ &= \sum_{p=h}^{\infty} \sum_{q=0}^{p} A_q B_{p-q} \vec{1} L^p r_t + \sum_{r=h}^{\infty} A_r L^{r-h} L^h e_t + G_r(r_t, L, lag < h) + G_e(e_t, L, lag < h) \end{aligned} \tag{7}$$

where $G_r(r_t, L, lag < h)$ and $G_e(e_t, L, lag < h)$ are the terms with r_{t-h+1} to r_t and e_{t-h+1} to e_t.
Therefore, the p_{t-h} as:

$$\Delta p_{t-h} = \vec{1} L^{h+1} r_t + \sum_{p=0}^{\infty} \sum_{q=0}^{p} A_q B_{p-q} \vec{1} r_t L^h L^p + \sum_{r=h}^{\infty} A_r L^r L^h e_t \tag{8}$$

Then $\mathrm{E}[\Delta p_t \Delta p'_{t-h}]$ can be written as following:

$$\begin{aligned} E[\Delta p_t \Delta p_{t-h}] &= \sum_{q=0}^{h+1} A_q B_{h+1-q} \vec{1}\vec{1}' \sigma^2_{rt} + \sum_{p=h}^{\infty} (\sum_{q=0}^{p} A_q B_{p-q}) \vec{1}\vec{1}' (\sum_{q=0}^{p-h} A_q B_{p-h-q})'] \sigma^2_{r_t} + \\ &\quad \sum_{r=h}^{\infty} A_r \Omega A'_{p-h} \end{aligned} \tag{9}$$

where $\sigma^2_{r_t}$ is the variance of efficient price shock r_t, the Ω is the variance-covariance matrix of the transitory noise e_t.
The equation(9) are our constraints for all h $\in \mathbb{N}$. Therefore, our general model captures all second moment restrictions across all lags. For the infinity equation sets equation(9) for h in [0,∞), in general we can say that in many cases, this is strong enough to identify/over-identify the parameter set $[A_p, B_q, \sigma^2_{r_t}, \Omega]$ for multivariate model. If not, we can easily add some restrictions (e.g., De Jong and Schotman, 2010, who is restricting the transitory noise variance-covariance matrix diagonal) for identification. This is because A_p are constructed by the

series of $\phi_{p,1}$, and the $\phi_{p,1}$ is set from i coefficient metrics $\Phi_{p_q 1}$ from equation(5), and B_q are constructed by the finite decomposition series of $\psi_{q,2}$. Therefore, in general we have infinity amount of restrictions from equation(9) by infinity amount of h from $[0, \infty)$, and we are estimating i K-by-K coefficient matrix A_p, n K-by-K coefficient matrix of B_q, a scalar $\sigma^2_{r_t}$ and a K-by-K variance-covariance matrix Ω.

We are aware that we did not provide strict analytical proof of when and why equation(9) is strong enough to identify/over-identify the coefficients because (1) we believe this is far beyond our scope (2) for most of the use cases, this can be easily identified by following the special cases we provide, and (3) for the models with too many lagging terms it might not be realistic because of the extremely heavy computational burden. We provide an example of how the special case of only lag 1 of endogenous and cointegration error-correction and over/under-reaction of the efficient price shock model is identified, we believe this should be the good enough because (1) reasonable computational burden (2) it already captures both of the error-correction mechanisms with further lagging (these terms are gradually decayed in an exponential pattern). Before providing the identification of special cases, we would first argue why the two error-correction mechanisms of lagging 1 is already reasonably capturing the behavior of further lagging.

1.2 Two Error-Correction Mechanism Analysis

In this part, by discussing two markets model with lag 1 error-correction terms[4], we argue that the lag 1 terms has already captured the further lagging error-correction mechanisms with reasonable structure. In other words, we discuss what endogenous error-correction mechanism really captures in the multivariate case.

For simplicity, we consider this simplified model with two markets (s_t and f_t) with over/under-reaction to the efficient price shock and only include lagging 1 of error-correction mechanisms:

$$\begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} s_{t-1} \\ f_{t-1} \end{pmatrix} + \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \tag{10}$$

Which is equivalent to the following permanent-transitory decomposition:

$$\begin{aligned} \begin{pmatrix} s_t \\ f_t \end{pmatrix} = \begin{pmatrix} m_t \\ m_t \end{pmatrix} + \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} (m_t - m_{t-1}) - \begin{pmatrix} \alpha_{11} + \alpha_{12} - 1 & 0 \\ 0 & \alpha_{21} + \alpha_{22} - 1 \end{pmatrix} \\ \begin{pmatrix} s_{t-1} - m_{t-1} \\ f_{t-1} - m_{t-1} \end{pmatrix} + \begin{pmatrix} \alpha_{12} \times (s_{t-1} - f_{t-1}) \\ \alpha_{21} \times (f_{t-1} - s_{t-1}) \end{pmatrix} + \begin{pmatrix} es_t \\ ef_t \end{pmatrix} \end{aligned} \tag{11}$$

[4]Error-correction terms of t-1, e.g. $s_{t-1} - m_{t-1}$.

For simplicity, we rewrite the model by setting the following:

$$\begin{aligned} G &= \begin{pmatrix} \gamma_s - 1 \\ \gamma_f - 1 \end{pmatrix} \\ M &= -\begin{pmatrix} \alpha_{11} + \alpha_{12} - 1 & 0 \\ 0 & \alpha_{21} + \alpha_{22} - 1 \end{pmatrix} \\ N &= \begin{pmatrix} \alpha_{12} & 0 \\ 0 & -\alpha_{21} \end{pmatrix} \\ t_t &= \begin{pmatrix} s_t - m_t \\ f_t - m_t \end{pmatrix} \\ c_t &= s_t - f_t \\ e_t^\mu &= G(m_t - m_{t-1}) + e_t \end{aligned} \tag{12}$$

Therefore, the model can be written as: [5]

$$\begin{aligned} t_t &= Mt_{t-1} + Nc_{t-1} + e_t^\mu \\ (I - ML)t_t &= NLc_t + e_t^\mu \\ t_t &= (I - ML)^{-1}NLc_t + (I - ML)^{-1}e_t^\mu \\ t_t &= (I + ML + M^2L^2 + ...)NLc_t + (I + ML + M^2L^2 + ...)e_t^\mu \\ t_t &= (\sum_{i=0}^{\infty}(ML)^i)Nc_{t-1} + (\sum_{i=0}^{\infty}(ML)^i)e_t^\mu \end{aligned} \tag{13}$$

Or, equivalently, if we consider finite length time series, we can get same relationship by iteration.

$$\begin{aligned} t_t &= Mt_{t-1} + Nc_{t-1} + e_t^\mu \\ &= M(Mt_{t-2} + Nc_{t-2} + e_{t-1}^\mu) + Nc_{t-1} + e_t^\mu \\ &= M^2t_{t-2} + N(c_{t-1} + Mc_{t-2}) + e_t^\mu + Me_{t-1}^\mu \\ &... \\ &= M^{t-1}t_1 + N\sum_{i=1}^{t-1}M^{i-1}c_{t-i} + \sum_{i=1}^{t-1}M^{i-1}e_{t-i+1}^\mu \end{aligned} \tag{14}$$

From these two representations above, we can clearly see that the t_t can be decomposed into the weighted sum of lagged cointegration error-correction term c_t, lagged efficient price change $m_t - m_{t-1}$ and noise term, till the time of t-1 (not including t-1).
Therefore, the endogenous error-correction terms of t_{t-1} is the linear combination of the lagged terms before t-1. In another word, our model have three patterns: endogenous error-correction terms (e.g. $s_{t-1} - m_{t-1}$) is responsible for all information before t-1. Specifically, it takes over the over/under-reaction of the cointegration error-correction term, efficient price change, and market

[5] We still assume $||M|| < 1$.

microstructure noise before t-1 with a decaying speed of matrix M[6]. This is in line with Andersen et al. 2022 paper, indicating their error-correction term is capturing the past transitory part before t-1. This is also reflected in our special case identification discussion in the following sections. The cointegration error-correction term captures the price level difference of these series at time t-1. The efficient price change term captures the over/under-reaction of efficient price change at time t-1. Therefore, in general, these three terms capture totally different patterns, which makes them orthogonal from one to the other.
Back to our original question, clearly from equation(14) we can see that the endogenous error-correction term captures the past endogenous pricing-error, past cointegration pricing-error, and past over/under-reaction of efficient price shock, with a reasonable decay rate of matrix M. Therefore, even our special case model only included the lag 1 of the error-correction terms in the representation, the slow decay patterns of further past pricing-errors are already captured [7]. Hence, we would consider the special case model that only includes lag 1 of the error-correction terms.

1.3 Identification analysis

Due to the complexity of the identification analysis of the most generalized framework, also since the lag 1 endogenous error-correction terms reasonable well capturing the further past two error-correction patterns with acceptable decay rate, we discuss the identification issue of this special case model with over/under-reaction of efficient price shock and lag 1 of error-correction mechanisms. Moreover, this identification analysis process can also be easily applied for most of the special cases of our generalized framework. We consider the model below:

$$p_t = \vec{1}m_t + \alpha r_t + \Phi(p_{t-1} - \vec{1}m_{t-1}) + e_t \tag{15}$$

For the K market case, p_t is a K×1 array of asset price, $\vec{1}$ is a K×1 vector of 1, m_t is the efficient price scalar, Φ is a K×K matrix (Notice, since we already fixed the lagging structure, Φ is a matrix but not a function), α is a K×1 array of coefficients, and e_t is a K×1 array of noise.
Since we are assuming the noise follows normal distribution, and following De Jong and Schotman (2010) and various literature, we calculate the second moment conditions as constraints. But first, we make some preparations from equation(15) (where I is the K×K identity matrix):

$$\begin{aligned} p_t - \vec{1}m_t &= \Phi L(p_t - \vec{1}m_t) + \alpha r_t + e_t \\ (I - \Phi L)(p_t - \vec{1}m_t) &= \alpha r_t + e_t \\ (p_t - \vec{1}m_t) &= (I - \Phi L)^{-1}(\alpha r_t + e_t) \end{aligned} \tag{16}$$

[6] $||M|| < 1$.

[7] As we are expecting, the further lagged pricing-errors should be less significantly affecting future price change comparing to the more recent pricing-errors pattern. In other words, investors would care more about more recent price series rather than the further lagged ones.

Set $(I-\Phi L)^{-1}$=$\Gamma(L)$,[8] the price change Δp_t calculated as following:

$$\begin{aligned}
p_t &= \vec{1}m_t + \alpha r_t + \Phi(p_{t-1} - \vec{1}m_{t-1}) + e_t \\
p_t - \vec{1}m_t - \Phi L(p_t - \vec{1}m_t) &= \alpha r_t + e_t \\
(I-\Phi L)(p_t - \vec{1}m_t) &= \alpha r_t + e_t \\
p_t - \vec{1}m_t &= (I-\Phi L)^{-1}(\alpha r_t + e_t) \\
p_t &= \vec{1}m_t + (I-\Phi L)^{-1}(\alpha r_t + e_t) \\
\Delta p_t &= \vec{1}r_t + (I-L)(I-\Phi L)^{-1}(\alpha r_t + e_t)
\end{aligned} \tag{17}$$

Since $(I-\Phi L)^{-1} = I + \Phi L + \Phi^2 L^2 + ...$ when $||\Phi|| < 1$.

$$\begin{aligned}
\Delta p_t &= \vec{1}r_t + (I-L)(I + \Phi L + \Phi^2 L^2 + ...)(\alpha r_t + e_t) \\
&= \vec{1}r_t + [I + (\Phi - I)L + \Phi(\Phi - I)L^2 + \Phi^2(\Phi - I)L^3 + ...](\alpha r_t + e_t) \\
&= (\vec{1} + \alpha)r_t + e_t + (\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t)
\end{aligned} \tag{18}$$

With the representation above, we can write out the second moment constraints.

$$\begin{aligned}
E[\Delta p_t \Delta p_t'] &= E[((\vec{1} + \alpha)r_t + e_t + (\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t)) \\
&\quad ((\vec{1} + \alpha)r_t + e_t + (\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t))'] \\
&\quad = E[((\vec{1} + \alpha)r_t + e_t + (\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t)) \\
&\quad ((\vec{1} + \alpha)'r_t + e_t' + (\alpha r_t + e_t)'(L + \Phi L^2 + \Phi^2 L^3 + ...)'(\Phi - I)'] \\
&\quad = E[\vec{1}\vec{1}r_t^2 + \vec{1}\alpha' r_t^2 + \alpha\vec{1}' r_t^2 + (\alpha r_t + e_t)(\alpha r_t + e_t)' + (\Phi - I) \\
&\quad (L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t)(\alpha r_t + e_t)'(L + \Phi L^2 + \Phi^2 L^3 + ...)' \\
&\quad (\Phi - I)'] \\
&\quad = \vec{1}\vec{1}\sigma_{rt}^2 + \vec{1}\alpha'\sigma_{rt}^2 + \alpha\vec{1}'\sigma_{rt}^2 + \Sigma + (\Phi - I) \\
&\quad E[(L(\alpha r_t + e_t) + \Phi L^2(\alpha r_t + e_t) + ...)(L(\alpha r_t + e_t)' + L^2(\alpha r_t + e_t)'\Phi' + ...)] \\
&\quad (\Phi - I)'] \\
&\quad = \vec{1}\vec{1}\sigma_{rt}^2 + \vec{1}\alpha'\sigma_{rt}^2 + \alpha\vec{1}'\sigma_{rt}^2 + \Sigma + (\Phi - I)(\Sigma + \Phi\Sigma\Phi_T + \Phi^2\Sigma\Phi_t^2 + ...) \\
&\quad (\Phi_T - I) \\
&\quad = \vec{1}\vec{1}\sigma_{rt}^2 + \vec{1}\alpha'\sigma_{rt}^2 + \alpha\vec{1}'\sigma_{rt}^2 + \Sigma + \sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I)
\end{aligned} \tag{19}$$

where $\Sigma = E[(\alpha r_t + e_t)(\alpha r_t + e_t)']$ and $\Phi_T = \Phi'$. The equation above is simplified because $cov(r_i, e_j) = 0$ for all [i,j] in $\mathbb{N}$, and $cov(r_i, r_j) = cov(e_i, e_j) = 0$ for all[i,j] in $\mathbb{N}$ and $i \neq j$. Therefore, all cross terms between r_t and e_t from

[8]When $||\Phi|| < 1$, the equation $\Gamma(L) = I + \Phi L + \Phi^2 L^2 + ...$ feasible because Φ can be decomposed by spectral decomposition, $\Phi = A\Lambda A'$, where is the diagonal matrix with eigenvalues. Therefore, Φ^k converges to 0 when K is infinity, since $\Phi^k = A\Lambda^k A'$ and all absolute eigenvalues smaller than 1 and converge to 0.

different lagging terms can be ignored.
Similarly, we can also calculate the lagged moment constraints as following:

$$\begin{aligned}
E[\Delta p_t \Delta p'_{t-h}] &= E[(\vec{1}+\alpha)r_t + e_t + (\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t) \\
&\quad (L^h(\vec{1}+\alpha)r_t + L^h e_t + L^h(\Phi - I)(L + \Phi L^2 + \Phi^2 L^3 + ...)(\alpha r_t + e_t))'] \\
&= E[((\vec{1}+\alpha)r_t + e_t + (\Phi - I)(L(\alpha r_t + e_t) + \Phi L^2(\alpha r_t + e_t) + ...)) \\
&\quad L^h((\vec{1}+\alpha)r_t + e_t + (\Phi - I)(L(\alpha r_t + e_t) + \Phi L^2(\alpha r_t + e_t) + ...))'] \\
&= E[((\Phi - I)(\Phi^{h-1}L^h(\alpha r_t + e_t) + \Phi^h L^{h+1}(\alpha r_t + e_t) + \Phi^{h+1}L^{h+2}(\alpha r_t + e_t) \\
&\quad + ...))L^h((\vec{1}+\alpha)r_t + e_t + (\Phi - I)(L(\alpha r_t + e_t) + \Phi L^2(\alpha r_t + e_t) + ...))'] \\
&= (\Phi - I)\Phi^{h-1}E[(\alpha r_t + e_t)((\vec{1}+\alpha)r_t + e_t)' + \Phi(\alpha r_t + e_t)(\alpha r_t + e_t)' \\
&\quad (\Phi_T - I) + \Phi^2(\alpha r_t + e_t)(\alpha r_t + e_t)'\Phi_T(\Phi_T - I)] \\
&= (\Phi - I)\Phi^{h-1}\alpha\vec{1}'\sigma_{rt}^2 + (\Phi - I)\Phi^{h-1}\Sigma + (\Phi - I)\Phi^{h-1}\Phi\Sigma(\Phi_T - I) + \\
&\quad (\Phi - I)\Phi^{h-1}\Phi^2\Sigma\Phi_T(\Phi_T - I) + ... \\
&= \Phi^{h-1}[(\Phi - I)\alpha\vec{1}'\sigma_{rt}^2 + (\Phi - I)\Sigma + (\Phi - I)\Phi\Sigma(\Phi_T - I) + \\
&\quad (\Phi - I)\Phi^2\Sigma\Phi_T(\Phi_T - I) + ...] \\
&= \Phi^{h-1}[(\Phi - I)\alpha\vec{1}'\sigma_{rt}^2 + (\Phi - I)\Sigma + \Phi\sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I)]
\end{aligned} \tag{20}$$

From the equation(20), we can see that $(\Phi - I)\alpha\vec{1}'\sigma_{rt}^2 + (\Phi - I)\Sigma + \Phi\sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I)$ is unrelated with h. Therefore, by changing h from 1 to ∞, we can have these moment constraints to over identify Ψ, from $E(\Delta p_t \Delta P'_{t-h})$ as follows:

$$\begin{aligned}
&E[\Delta p_t \Delta p'_{t-2}] = \Phi E[\Delta p_t \Delta p'_{t-1}] \\
&E[\Delta p_t \Delta p'_{t-3}] = \Phi^2 E[\Delta p_t \Delta p'_{t-1}] \\
&E[\Delta p_t \Delta p'_{t-4}] = \Phi^3 E[\Delta p_t \Delta p'_{t-1}] \\
&...
\end{aligned} \tag{21}$$

Therefore, the coefficient matrix Φ can be over-identified from these constraints. Then for the next step, to identify the rest parameters (Σ,α and σ_{rt}^2), we have these two constraint matrices:

$$\begin{aligned}
C1 &= \vec{1}\vec{1}\sigma_{rt}^2 + \vec{1}\alpha'\sigma_{rt}^2 + \alpha\vec{1}'\sigma_{rt}^2 + \Sigma + \sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I) \\
C2 &= (\Phi - I)\alpha\vec{1}'\sigma_{rt}^2 + (\Phi - I)\Sigma + \Phi\sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I)
\end{aligned} \tag{22}$$

where C1 and C2 are two constant metrics constraints, we can get $C1 = E[\Delta p_t \Delta p'_t]$ and $C2 = \Phi^{-h+1}E[\Delta p_t \Delta p'_{t-h}]$.

Generally speaking, there are K(K+1)/2 constraints from C1, since both right and left side of equation(19) are symmetry matrices and there are K^2 constraints since the right and left side of equation(20) are asymmetry matrices.
Therefore, we vectorize the equation(19) and equation(20) to consider the identification of the model:

$$\begin{aligned} Vec(C1) &= Vec[(\vec{1}+\alpha)(\vec{1}+\alpha)'\sigma_{rt}^2] + Vec(\Omega) + Vec[\sum_{k=0}^{\infty}(\Phi - I)\Sigma\Phi_T^k(\Phi_T - I)] \\ &= Vec(\vec{1}\vec{1}' + \vec{1}\alpha' + \alpha\vec{1}')\sigma_{rt}^2 + Vec(\Sigma) + \sum_{k=0}^{\infty}[(\Phi - I)\Phi^k]\otimes \\ &[(\Phi - I)\Phi^k]Vec(\Sigma) \end{aligned} \tag{23}$$

To simplify the equation, we set $\Theta = \sum_{k=0}^{\infty}[(\Phi - I)\Phi^k] \otimes [(\Phi - I)\Phi^k]$, then the vectorized C1 is following:[9].

$$\begin{aligned} Vec(C1) &= \vec{1}I\vec{1}\sigma_{rt}^2 + I \otimes \vec{1}Vec(\alpha')\sigma_{rt}^2 + \vec{1} \otimes IVec(\alpha)\sigma_{rt}^2 + (I^{K^2} + \Theta)Vec(\Sigma) \\ &= \vec{1}_{K^2}\sigma_{rt}^2 + (I \otimes \vec{1} + \vec{1} \otimes I)\alpha\sigma_{rt}^2 + (I_{K^2} + \Theta)Vec(\Sigma) \end{aligned} \tag{24}$$

where $\vec{1}\otimes I\vec{1} = \vec{1}_{k^2}$, which is a matrix size of $K^2 \times 1$, I_{K^2} is the $K^2 \times K^2$ identity matrix.
Similarly, we also vectorize the constraints C2 as following:

$$\begin{aligned} Vec(C2) &= Vec[(\Phi - I)\alpha\vec{1}']\sigma_{rt}^2 + Vec[(\Phi - I)\Sigma I] + \\ &Vec[\Phi\sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I)I] \\ &= \vec{1} \otimes (\Phi - I)\alpha\sigma_{rt}^2 + I \otimes (\Phi - I)Vec(\Sigma) + I \otimes \Phi \\ &Vec(\sigma_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi - T - I)) \\ &= \vec{1} \otimes (\Phi - I)\alpha\sigma_{rt}^2 + I \otimes (\Phi - I)Vec(\Sigma) + I \otimes \Phi \\ &\sum_{k=0}^{\infty}[(\Phi - I)\Phi^k] \otimes [(\Phi - I)\Phi^k]Vec(\Sigma) \\ &= \vec{1} \otimes (\Phi - I)\alpha\sigma_{rt}^2 + [I \otimes (\Phi - I) + I \otimes \Phi\Theta]Vec(\Sigma) \end{aligned} \tag{25}$$

[9]Here we use $Vec(AXB) = (B' \otimes A)Vec(X)$ and $Vec(\alpha) = Vec(\alpha')$ since α is a vector.

Before we further discuss the identification of the model, we need to simplify Θ to simplify the identification constraints[10]:

$$\begin{aligned}\Theta &= \sum_{k=0}^{\infty}[(\Phi - I)\Phi^k] \otimes [(\Phi - I)\Phi^k] \\ &= \sum_{k=0}^{\infty}(\Phi - I) \otimes (\Phi - I)(\Phi^k \otimes \Phi^k) \\ &= (\Phi - I) \otimes (\Phi - I)(sum_{k=0}^{\infty}(\Phi \otimes \Phi)) \\ &= (\Phi \otimes \Phi + I \otimes I - \Phi \otimes I - I \otimes \Phi)(I_{K^2} - \Phi \otimes \Phi)^{-1} \\ &= I_{K^2} + (2\Phi \otimes \Phi - \Phi \otimes I - I \otimes \Phi)(I_{K^2} - \Phi \otimes \Phi)^{-1}\end{aligned} \tag{26}$$

After splitting out the parameters (Σ, α and σ_{rt}^2) by vectorization, we have a clearer structure for identification issue. Generally speaking, the identification issue can be discussed by cancel out parameter through the union/combine of Vec(C1) and Vec(C2), since we have $K^2 + K(K+1)/2$ constraints to identify $K + K(K+1)/2 + 1$ parameters/coefficients. If there is identification issue, additional restriction can be added to the model. We aware this might be too general for the identification issue discussion. Therefore, we will provide an example of a two markets case below.

Corollary 1 Identification of two markets case (K=2)
For the K=2 case (two markets case), we examine the identification by canceling out α.
Set $C3 = Vec(C2) - \vec{1} \otimes (\Phi - I)N_L^+ Vec(C1)$, where N_L^+ is the Moore–Penrose left inverse of the matrix $(I \otimes \vec{1} + \vec{1} \otimes I)$.
Therefore, we can easily calculate C3 as follows:

$$\begin{aligned}C3 = &(I \otimes (\Phi - I) + (I + \Phi)\Theta)Vec(\Sigma) - \vec{1} \otimes (\Phi - I)N_L^+(I_{K^2} + \Theta)Vec(\Sigma) - \\ &\vec{1} \otimes (\Phi - I)N_L^+(I \otimes \vec{1})\vec{1}\sigma_{rt}^2 \\ &= [(I \otimes (\Phi - I) + (I \otimes \Phi)\Theta) - \vec{1} \otimes (\Phi - I)N_L^+(I_{K^2} + \Theta)]Vec(\Sigma) \\ &- \vec{1} \otimes (\Phi - I)N_L^+(I \otimes \vec{1})\vec{1}\sigma_{rt}^2\end{aligned} \tag{27}$$

Generally speaking, the equation above has 4 constraints (since C3 is a 4×1 matrix), and we need to identify 4 parameters (3 from Σ and 1 from σ_{rt}^2). However, we need to restrict the Σ diagonal so that the model is identifiable. Here we will indicate why we need this additional restriction and why when Σ is diagonal the model can be identified.
For the general case $C3 = PVec(\Sigma) + Q\sigma_{rt}^2 = [P1, P2, P3, P4]Vec(\Sigma) + [Q]\sigma_{rt}^2$,

[10] We are using $(A \otimes C)(B \otimes D) = (AC) \otimes (BD)$ and $(I - A)^{-1} = I + A + A^2 + A^3 + ...$, where $||A|| < 1$.

where Pi is the ith column of the matrix P.
The constraints as following:

$$\begin{aligned} C3 &= [P1, P2, P3, P4][Var(e_{st}), Cov(e_{st}, e_{ft}), Cov(e_{st}, e_{ft}), Var(e_{ft})]' + Q\sigma_{rt}^2 \\ &= [P1, P2 + P3, P4][Var(e_{st}, Cov(e_{st}, e_{ft}), Var(e_{ft})] + Q\sigma_{rt}^2 \end{aligned} \tag{28}$$

We can find the rank([P1,P2+P3,P4])=2.[11] Therefore, equation(27) is not strong enough to identify 3 parameters/coefficients, so that we need to restrict the Σ diagonal.
On the other hand, rank([Q,P1,P4])=3. Considering equation(27) is in the representation of Ax=b with 4 rows, Σ (diagonal) and σ_{rt}^2 can be (over)-identified. Specifically, equation(27) can be specified as equations below to identify σ_{rt}^2, $Var(e_{st})$, $Cov(e_{st}, e_{ft})$ and $Var(e_{ft})$:

$$\begin{aligned} q1 \times \sigma_{rt}^2 + p11 \times Var(e_{st}) + (p12 + p13) \times Cov(e_{st}, e_{ft}) + p14 \times Var(e_{ft}) = C31 \\ q2 \times \sigma_{rt}^2 + p21 \times Var(e_{st}) + (p22 + p23) \times Cov(e_{st}, e_{ft}) + p24 \times Var(e_{ft}) = C32 \\ q3 \times \sigma_{rt}^2 + p31 \times Var(e_{st}) + (p32 + p33) \times Cov(e_{st}, e_{ft}) + p34 \times Var(e_{ft}) = C33 \\ q4 \times \sigma_{rt}^2 + p41 \times Var(e_{st}) + (p42 + p43) \times Cov(e_{st}, e_{ft}) + p44 \times Var(e_{ft}) = C34 \end{aligned} \tag{29}$$

where qi is the ith row of Q, Pij is the ith row of Pi, C3i is the ith row of C3.
For the equations above, all four parameters can be identified if and only if [Q,P1,P2+P3,P4] is full rank. But we can also identify the rank of [Q,P1,P2+P3,P4] is 3. Therefore, all these 4 parameters (σ_{rt}^2, $Var(e_{st})$, $Cov(e_{st}, e_{ft})$ and $Var(e_{ft})$) cannot be identified together.
On the other hand, if we restrict the Σ to be diagonal, we will identify σ_{rt}^2, $Var(e_{st})$ and $Var(e_{ft})$ from the equations below:

$$\begin{aligned} q1 \times \sigma_{rt}^2 + p11 \times Var(e_{st}) + p14 \times Var(e_{ft}) = C31 \\ q2 \times \sigma_{rt}^2 + p21 \times Var(e_{st}) + p24 \times Var(e_{ft}) = C32 \\ q3 \times \sigma_{rt}^2 + p31 \times Var(e_{st}) + p34 \times Var(e_{ft}) = C33 \\ q4 \times \sigma_{rt}^2 + p41 \times Var(e_{st}) + p44 \times Var(e_{ft}) = C34 \end{aligned} \tag{30}$$

Since rank([Q,P1,P4])=3, we can (over)identify σ_{rt}^2, $Var(e_{st})$ and $Var(e_{ft})$ from the equation above.
After we identify Σ and σ_{rt}^2, we can identify α from equation(19) as follows:

$$C1 = \vec{1}\vec{1}\sigma_{rt}^2 + \vec{1}\alpha'\sigma_{rt}^2 + \alpha\vec{1}'\sigma_{rt}^2 + \Sigma + \sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I) \tag{31}$$

We set $C4 = [C1 - \Sigma - \sum_{k=0}^{\infty}(\Phi - I)\Phi^k\Sigma\Phi_T^k(\Phi_T - I) - \vec{1}\vec{1}'\sigma_{rt}^2](\sigma_{rt}^2)^{-1}$. The constant matrix C4 has 3 constraints to identify α. If we set $\alpha = [a, b]'$. The

[11] Due to the space limitation, we won't show the calculation here. But this can be easily checked by calculating software such as Matlab.

identification can be simplified as follows:

$$\begin{aligned} C4 &= \vec{1}\alpha' + \alpha\vec{1}' \\ &= \begin{pmatrix} 2a & a+b \\ a+b & 1 \end{pmatrix} \end{aligned} \tag{32}$$

Therefore, [a,b] or α can be identified.

2 Information share framework

In this part, we provide the steps to calculate the information share for the generalized model following Hasbrouck and De Jong and Schotman's idea[12], and this analysis is based on the general case:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \Phi_L(p_t, m_t) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{33}$$

Where m_t is the scalar of efficient price, r_t is the scalar of efficient price change, since we are modeling a K market case, p_t is a K $\times$ 1 vector, α is a K $\times$ 1 matrix, I is a K $\times$ K identity matrix, $\vec{1}$ is a K $\times$ 1 vector of 1, Φ_L is a function of matrix with lagged p_t and m_t with $Cov(r_t, r_{t-i}) = 0$, $Cov(e_t, e_{t-i}) = 0$, $Cov(r_t, e_t) = 0$ and $Cov(r_t, e_{t-i}) = 0$ for all i $\in \mathbb{N}$.
In the general case, we still decompose $\Phi_L(p_t, m_t)$ as $\Phi_{1L}(p_t - \vec{1}m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1}$, where Φ_{1L} and Φ_{2L} are polynomial functions of a matrix with the lagging indicator.
Therefore, the model can be rewritten as follows by separating the generalized error-correction terms:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \Phi_{1L}(p_t - \vec{1}m_{t-1}) - \Phi_{2L}(m_t - m_{t-1})\vec{1} + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{34}$$

Following De Jong and Schotman (2010), we define the price innovations as $v_t = p_t - \vec{1}m_{t-1}$.
Therefore, the model can be written as:

$$\begin{aligned} (I - \Phi_{1L})v_t &= (\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + e_t \\ v_t &= (I - \Phi_{1L})^{-1}(\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + (I - \Phi_{1L})^{-1}e_t \end{aligned} \tag{35}$$

In this case, we assume $||\Phi_{1L}|| < 1$ thus $(I - \Phi_{1L})^{-1} = \phi_{0,1} + \phi_{1,1}L + \phi_{2,1}L^2 + ...$, and $\Phi_{2L} = \psi_{0,2} + \psi_{1,2}L + \psi_{2,2}L^2 + ...\psi_{n,2}L^n$. As mentioned, this representation is normally achieved by unfolding following decomposition to factorization from the matrix polynomial of L to the product of the highest power of 1. Therefore,

[12]Since the α coefficient indicate the over/under-reaction to the efficient price shock, we leave this in the equation for all estimations besides the identification analysis.

we can move the lagging indicators out and Ψ_{i1} matrices no longer include lagging indicator L.

$$\begin{aligned}(I-\Phi_{1L})^{-1} &= \phi_{01}(I-\Phi_{11}L)^{-1}(I-\Phi_{21}L)^{-1}...(I-\Phi_{i1}L)^{-1} \\ &= \phi_{01}(I+\Phi_{11}L+\Phi_{11}^2L^2+...)(I+\Phi_{21}L+\Phi_{21}^2L^2+...)... \\ &(I+\Phi_{i1}L+\Phi_{i1}^2L^2+...) \\ &= \phi_{01}\prod_{k=1}^{i}\sum_{j=1}^{\infty}(I+\Phi_{k1}^jL^j)\end{aligned} \tag{36}$$

Therefore, the general model or equation(35) can be modified as following:

$$\begin{aligned}v_t = (\phi_{0,1}+\phi_{1,1}L+\phi_{2,1}L^2+...)(\vec{1}+\alpha-(\psi_{0,2}+\psi_{1,2}L+\psi_{2,2}L^2+...\psi_{n,2}L^n)\vec{1})r_t \\ +(\phi_{0,1}+\phi_{1,1}L+\phi_{2,1}L^2+...)e_t\end{aligned} \tag{37}$$

All lagged r_t and all e_t are uncorrelated with r_t by assumption. Since we are following De Jong and Schotman's information share representation, the IS is based on the regression of r_t to v_t (where $v_t = p_t - m_{t-1}$), so it is straightforward to simplify eq.(35) as following:

$$v_t = \phi_{0,1}(\vec{1}+\alpha-\psi_{0,2}\vec{1})r_t + \Phi_{other,L}L\vec{1}r_t + \sum_{i=0}^{\infty}\phi_{i,1}L^ie_t \tag{38}$$

Then following De Jong and Schotman, 2010, we run the linear regression $r_t = \gamma_{ols}v_t + \eta_t$, γ can be calculated as following:

$$\begin{aligned}\gamma &= \Upsilon^{-1}Cov(v_t, r_t) \\ &= \Upsilon^{-1}\phi_{01}(\vec{1}+\alpha-\psi_{0,2}\vec{1})\sigma^2\end{aligned} \tag{39}$$

Then the information share can be calculated as:

$$\begin{aligned}R^2 &= 1-\frac{\sigma_\eta^2}{\sigma^2} \\ &= \frac{\gamma'\Upsilon\gamma}{\sigma^2} \\ &= \gamma'[\phi_{01}(\vec{1}+\alpha-\psi_{0,2}\vec{1})]\end{aligned} \tag{40}$$

Therefore, the information share can be represented as follows

$$\begin{aligned}\gamma &= \Upsilon^{-1}\phi_{0,1}(\vec{1}+\alpha-\psi_{0,2}\vec{1})\sigma^2 \\ IS &= \gamma\circ\phi_{0,1}(\vec{1}+\alpha-\psi_{0,2}\vec{1}). \\ \Upsilon &= E[v_tv_t']\end{aligned} \tag{41}$$

The equation(41) is the generalized IS estimator from De jong and Schotman, where they restricted $\phi_{0,1} = I$ and $\psi_{0,2} = 0$[13]. By the generalized representation equation(1), the framework would easily fit with various multivariate market microstructure models.

2.1 Variance-covariance matrix of v_t

On the other hand, to calculate the IS from equation(41), we need to provide solution of Υ^{-1}. Here we indicate how to get the expression of Υ^{-1} following Lutkepohl (2005), Chapter 11.

We start the calculation from equation(35) by decomposing $\Phi_{1L} = \psi_{0,1}+\psi_{1,1}L+\psi_{2,1}L^2+...\psi_{n,1}L^n$:

$$\begin{aligned}
&(I - \Phi_{1L})v_t = (\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + e_t \\
&(\psi_{0,1} + \psi_{1,1}L + \psi_{2,1}L^2 + ...\psi_{n,1}L^n)v_t = (\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + e_t \\
&\psi_{0,1}v_t = -(\psi_{1,1}L + \psi_{2,1}L^2 + ...\psi_{n,1}L^n)v_t + (\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + e_t \\
&v_t = -\psi_{0,1}^{-1}(\psi_{1,1}L + \psi_{2,1}L^2 + ...\psi_{n,1}L^n)v_t + \psi_{0,1}^{-1}(\vec{1} + \alpha - \Phi_{2L}\vec{1})r_t + \psi_{0,1}^{-1}e_t
\end{aligned} \tag{42}$$

To simplify the equation above, we set:

$$\begin{aligned}
A_i &= -\psi_{0,1}^{-1}\psi_{i,1} \\
\psi_{0,1}^{-1}(\vec{1} + \alpha - \Phi_{2L}\vec{1}) &= \sum_{i=0}^{q} B_i L^i \\
\psi_{0,1}^{-1}e_t &= u_t
\end{aligned} \tag{43}$$

Therefore, eq.(42) can be simplified as:

$$v_t = A_1 v_{t-1} + A_2 v_{t-2} + ... + A_p v_{t-p} + B_0 r_t + B_1 r_{t-1} + ... + B_q r_{t-q} + u_t \tag{44}$$

[13]We aware that they mention that the transitory noise in their model can be non-diagonal. However, in all the models, they included additional restrictions in their identification analysis, e.g., $Cov(e_t, r_t) = 0$ and BN decomposition, and restrict orthogonal e_t to restrict their IS ranging from [0,1].

Then this VAR(p)-X model can be rewritten as VAR(1) model as follows:

$$
\begin{aligned}
&Y_t = AY_{t-1} + U_t \\
&Y_t = \begin{pmatrix} v_t \\ v_{t-1} \\ ... \\ v_{t-p+1} \\ r_t \\ r_{t-1} \\ ... \\ r_{t-q+1} \end{pmatrix}; A = \begin{pmatrix} A_{11}, A_{12} \\ A_{21}, A_{22} \end{pmatrix}; U_t = \begin{pmatrix} u_t + B_0 r_t \\ 0 \\ 0 \\ ... \\ 0 \\ r_t \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix} \\
&A_{11} = \begin{pmatrix} A_1 & A_2 & ... & A_{p-1} & A_q \\ I_m & 0 & ... & 0 & 0 \\ 0 & I_m & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & I_m & 0 \\ 0 & 0 & ... & 0 & I_m \end{pmatrix}; A_{12} = \begin{pmatrix} B_1 & ... & B_q \\ 0 & ... & 0 \\ ... & ... & ... \\ 0 & ... & 0 \end{pmatrix} \\
&A_{21} = 0; A_{22} = \begin{pmatrix} 0 & 0 & ... & 0 & 0 \\ 1 & 0 & ... & 0 & 0 \\ 0 & 1 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & ... & 1 & 0 \end{pmatrix}
\end{aligned} \tag{45}
$$

where I_m is the $K \times K$ identity matrix (K is the length of v_t), and for U_t, r_t is in the pK+1 row.
Therefore, we can calculate the second moment of Y_t from equation(45)) (define $\Gamma_Y(h) = E[Y_t Y'_{t-h}]$):

$$
\begin{aligned}
Y_t &= AY_{t-1} + U_t \\
E[Y_t Y_{t-h}] &= AE[Y_{t-1} Y'_{t-h}] + E[U_t Y'_{t-h}] \\
\Gamma_Y(h) &= A\Gamma_Y(h-1) + E[U_t Y'_{t-h}]
\end{aligned} \tag{46}
$$

Considering when $h > 0$, $E[U_t Y'_{t-h}] = 0$ since u_t and r_t is not contained with lagged v_t and r_t series.
Then we can easily get:

$$
\begin{aligned}
\Gamma_Y(h) &= A\Gamma_Y(h-1) \\
\Gamma_Y(0) &= A\Gamma_Y(-1) + E[U_t Y'_t] \\
&= A\Gamma_Y(1)' + E[U_t Y'_t] \\
&= A\Gamma_Y(0)A' + E[U_t Y'_t] \\
vec(\Gamma_Y(0)) &= vec(A\Gamma_Y(0)A') + vec(E[U_t Y'_t]) \\
vec(\Gamma_Y(0)) &= (A \otimes A) vec(\Gamma_Y(0)) + vec(E[U_t Y'_t]) \\
vec(\Gamma_Y(0)) &= (I_{m \times m} - A \otimes A)^{-1} vec(E[U_t Y'_t])
\end{aligned} \tag{47}
$$

With the expression of $E[U_t Y_t]$ we can get the vectored $\Gamma_Y(0)$, then $Cov(v_t) = \begin{pmatrix} I_m & 0 & ... & 0 \\ 0 & 0 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & 0 \end{pmatrix} \Gamma_Y(0))$

Therefore, since the expression of U_t and Y_t can be taken from equation(45), we can get the expression of $E[U_t Y_t']$ as

$$\begin{aligned} E[U_t Y_t'] &= \begin{pmatrix} u_t u_t' + B_0 B_0' r_t^2 & 0 & ... & B_0 r_t^2 & ... & 0 \\ 0 & 0 & ... & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ B_0' r_t^2 & 0 & ... & r_t \times r_t & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & ... & 0 \end{pmatrix} \\ &= \begin{pmatrix} \psi_{0,1}^{-1} \Sigma_{e_t} \psi_{0,1}^{-1\prime} & 0 & ... & B_0 \sigma^2 & ... & 0 \\ 0 & 0 & ... & 0 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ B_0' \sigma^2 & 0 & ... & \sigma^2 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & ... & 0 \end{pmatrix} \end{aligned} \tag{48}$$

Where $r_t \times r_t$ and σ^2 is on the $pm + 1$ column and row. Note, B_i for i $\in$ [0,q] are $K \times 1$ vectors here.

We understand that for most empirical models, the calculation of the variance-covariance matrix does not need such a complex process (vectorization and so on). However, we still provide the variance-covariance matrix for the general framework of equation(34) since even though the representation might be cumbersome, the process can be easily standardized. More importantly, this process indicates that the price innovation from the state-space model actually follows a VARMA process after solving the unobserved component model.

2.2 Special cases of general information share indicator

Since the information share representation for the generalized market microstructure model is still too generalized, in most cases, the information share indicator can be calculated with an easier process. Here we discuss two corollaries that indicate how to calculate the information share in special cases. Notice, in these corollaries, we might have stronger restrictions than the general case and our empirical analysis (e.g., we assume Cov(e_t) is diagonal).

Corollary 2: When $\Phi_L = \beta(p_{t-1} - \vec{1} m_{t-1})$ and β is a K $\times$ K diagonal matrix, the information share can be express as $IS = \gamma \circ (\vec{1} + \alpha)$. This is a case when we add the endogenous error correction term within the De Jong and

Schotman's model.

Proof: Similar to the previous generous case, this model can be written as follows:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \beta(p_{t-1} - \vec{1}m_{t-1}) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{49}$$

where p_t is a K $\times$ 1 vector, $\vec{1}$ is a column vector of 1, m_t is the efficient price as a number, α is a K $\times$ 1 vector, β is a K $\times$ K diagonal matrix, Var(r_t)=σ^2, Var(e_t)=Ω, e_t is a K $\times$ 1 vector and Ω is a K $\times$ K diagonal matrix (which we assume that $E[cov(e_t, e_{t-i})] = 0$ for i $\in \mathbb{N}$).
Similarly, if we assume $v_t = p_t - m_{t-1}$, this model can be expressed as following:

$$\begin{aligned} v_t &= (\vec{1} + \alpha)r_t + \beta v_{t-1} - \beta\vec{1}Lr_t + e_t \\ (I - \beta L)v_t &= (\vec{1} + \alpha - \beta\vec{1}L)r_t + e_t \\ v_t &= (I - \beta L)^{-1}(\vec{1} + \alpha - \beta\vec{1}L)r_t + (I - \beta L)^{-1}e_t \end{aligned} \tag{50}$$

Where L is the lag indicator, and I is a ranking K identity matrix.
Therefore, the express of Υ can be calculated as following:

$$\begin{aligned} \Upsilon &= E[v_t v_t'] \\ &= [(I - \beta L)^{-1}(\vec{1} + \alpha - \beta\vec{1}L)r_t + (I - \beta L)^{-1}e_t] \\ &[(I - \beta L)^{-1}(\vec{1} + \alpha - \beta\vec{1}L)r_t + (I - \beta L)^{-1}e_t]' \end{aligned} \tag{51}$$

To calculate Υ, we set diagonal of β as B, where $B = [b_1, b_2, ...b_n]'$ and all ($\mid bi \mid < 1$),$\alpha = [a_1, a_2, ...a_n]'$, $e_t = [e_{t,1}, e_{t,2}, ...e_{t,n}]$ and the diagonal vector of Ω as $\omega = [\omega_1, \omega_2, ...\omega_n]$.
Thus, we can write v_t as following:

$$\begin{aligned} v_t &= (I - \beta L)^{-1}(\vec{1} + \alpha - \beta\vec{1}L)r_t + (I - \beta L)^{-1}e_t \\ &= [(I + \beta L + \beta^2 L^2 + ...)(\vec{1} + \alpha - \beta\vec{1}L)r_t + (I - \beta L)^{-1}e_t] \\ &= [(\vec{1} + \alpha)r_t + (\beta\vec{1} + \beta\alpha - I\beta\vec{1})Lr_t + (\beta^2\vec{1} + \beta^2\alpha - \beta^2\vec{1})L^2 r_t + \\ &... + (I - \beta L)^{-1}e_t] \\ &= [(\vec{1} + \alpha)r_t + \beta\alpha Lr_t + \beta^2\alpha L^2 r_t + ... + Ie_t + \beta Le_t + \beta^2 L^2 e_t + ...] \\ &= \begin{pmatrix} (1 + a_1)r_t + b_1 a_1 r_{t-1} + b_1^2 a_1 r_{t-2} + ... + e_{t,1} + b_1 e_{t-1,1} + b_1^2 e_{t-2,1}... \\ (1 + a_2)r_t + b_2 a_2 r_{t-2} + b_2^2 a_2 r_{t-2} + ... + e_{t,2} + b_2 e_{t-1,2} + b_2^2 e_{t-2,2}... \\ ... \\ (1 + a_n)r_t + b_n a_n r_{t-1} + b_n^2 a_n r_{t-2} + ... + e_{t,n} + b_n e_{t-1,n} + b_n^2 e_{t-2,n}... \end{pmatrix} \end{aligned} \tag{52}$$

Therefore, we can easily simplify the $E[v_t v_t']$ by two parts.

$$\begin{pmatrix} \tau_{1,1} & \tau_{1,2} & ...\tau_{1,n} \\ \tau_{2,1} & \tau_{2,2} & ...\tau_{2,n} \\ ... & & \\ \tau_{n,1} & \tau_{n,2} & ...\tau_{n,n} \end{pmatrix} \tag{53}$$

For the diagonal terms $\tau_{i,i}$ considering $\mid b_i \mid < 1$, $E[cov(e_i, e_j)] = 0$ when $i \neq j$ $E[cov(e_i, r_j)] = 0$ for any i and j, and $cov(r_i, r_j) = 0$ for $i \neq j$, can be simplified as following:

$$\begin{aligned} \tau_{i,i} &= E[(1+a_i)^2 r_t^2 + b_i^2 a_i^2 r_{t-1}^2 + b_i^4 a_i^2 r_{t-2} + ... + e_{t,i}^2 + b_i^2 e_{t-1,i} + b_i^4 e_{t-2,i} + ...] \\ &= (1 + 2a_i + \frac{a_i^2}{1-b_i^2})\sigma^2 + \frac{\omega_i^2}{1-b_i^2} \end{aligned} \tag{54}$$

For the non-diagonal terms $\tau_{i,j}(i \neq j)$, similarly can be simplified as following:

$$\begin{aligned} \tau_{i,j} &= E[(1+a_i)(1+a_j) r_t^2 + b_i b_j a_i a_j r_{t-1}^2 + b_i^2 b_j^2 a_i a_j r_{t-2}^2 + ...] \\ &= (1 + a_i + a_j + \frac{a_i a_j}{1 - b_i b_j})\sigma^2 \end{aligned} \tag{55}$$

With the format of Υ, we can just follow our general case to derive the representation of information share.
The relation between innovation in the efficient price and the shocks to individual prices:

$$r_t = \gamma_{ols}' v_t + \eta_t \tag{56}$$

Therefore the regression coefficients γ are as follows:

$$\begin{aligned} \gamma_{ols} &= \Upsilon^{-1} \mathrm{Cov}(r_t, v_t) \\ &= \Upsilon^{-1} \mathrm{Cov}(r_t, (I - \beta L)^{-1}(\vec{1} + \alpha - \beta L) r_t + (I - \beta L)^{-1} e_t) \\ &= \Upsilon^{-1} \mathrm{Cov}(r_t, (I + \beta L + \beta L^2 + ...)(\vec{1} + \alpha - \beta L) r_t) \\ &= \Upsilon^{-1} \mathrm{Cov}(r_t, I(\vec{1} + \alpha) r_t) \\ &= \Upsilon^{-1}(\vec{1} + \alpha)\sigma^2 \end{aligned} \tag{57}$$

Thus, the total fraction of the variance in the efficient price innovation r_t explained by the vector price innovation is:

$$R^2 = 1 - \frac{\sigma_\eta^2}{\sigma^2} = \frac{\gamma_{ols}' \Upsilon \gamma_{ols}}{\sigma^2} = \gamma_{ols}'(\vec{1} + \alpha) \tag{58}$$

Thus, the vector of information share (IS) can be defined following De Jong and Schotman:

$$IS = \gamma_{ols} \circ (\vec{1} + \alpha) \tag{59}$$

This is a special case of De Jong and Schotman's representation when additional endogenous endogenous error-correction terms exist in the transitory part.

Corollary 3: When Φ_{1L} does not including lagging indicator, there is a structural/instantaneous error correction term. For this case we set: $\Phi_{1L} = \phi_{01}$ and $\Phi_{2L} = 0$, therefore p_t is allowed within the error correction term.

Proof: the model can be written as:

$$\begin{aligned} p_t &= \vec{1}m_t + \alpha r_t + \phi_{01}(p_t - \vec{1}m_{t-1}) + e_t \\ m_t &= m_{t-1} + r_t \end{aligned} \tag{60}$$

The derivation as the previous part:

$$\begin{aligned} v_t &= (\vec{1} + \alpha)r_t + \phi_{01}v_t + e_t \\ (I - \phi_{01})v_t &= (\vec{1} + \alpha)r_t + e_t \\ v_t &= (I - \phi_{01})^{-1}(\vec{1} + \alpha)r_t + (I - \phi_{01})^{-1}e_t \end{aligned} \tag{61}$$

We need to assume the matrix $I - \phi_{01}$ is full rank. To simplify the writing in the function, we set $\psi_{01} = (I - \phi_{01})^{-1}$.
Therefore, the equation above can be simplified as follows:

$$v_t = \psi_{01}(\vec{1} + \alpha)r_t + \psi_{01}e_t \tag{62}$$

Similar as before, when running the regression $r_t = \gamma_{ols}v_t + \eta_t$, γ_{ols} can be calculated as following:

$$\begin{aligned} \gamma_{ols} &= \Upsilon^{-1}Cov(v_t, r_t) \\ &= \Upsilon^{-1}\psi_{01}(\vec{1} + \alpha)\sigma^2 \end{aligned} \tag{63}$$

Then the information share can be calculated as before:

$$\begin{aligned} R^2 &= \frac{\gamma'_{ols}\Upsilon\gamma_{ols}}{\sigma^2} \\ &= \gamma'_{ols}[\psi_{01}(\vec{1} + \alpha)] \end{aligned} \tag{64}$$

$$IS = \gamma_{ols} \circ \psi_{01}(\vec{1} + \alpha) \tag{65}$$

Then there is the calculation of Υ:

$$\begin{aligned} \Upsilon &= E(v_t v'_t) \\ &= E(\psi_{01}(\vec{1} + \alpha)r_t + \psi_{01}e_t)(\psi_{01}(\vec{1} + \alpha)r_t + \psi_{01}e_t)' \\ &= E(\psi_{01}(\vec{1} + \alpha)r_t + \psi_{01}e_t)((\vec{1} + \alpha)'\psi'_{01}r_t + e'_t\psi'_{01}) \\ &= \psi_{01}(\vec{1} + \alpha)(\vec{1} + \alpha)'\psi'_{01}\sigma^2 + \psi_{01}\Omega\psi'_{01} \end{aligned} \tag{66}$$

3 Comparing generalized information share and De Jong and Schotman's information share

De Jong and Schotman's information share model uses the diagonal restriction of the variance-covariance matrix of transitory term e_t throughout their study. Thus the asset price is only connected by the r_t term (both permanent and transitory parts). However, we remove this restriction to allow additional co-movement across different markets. Therefore, the asset prices (p_t or v_t) is connect not only by the term of r_t, but also by the lagged r_t and e_t terms. This is clearly indicated from the equation(35): $v_t = (I-\phi_{01})^{-1}(\vec{1}+\alpha)r_t+(I-\phi_{01})^{-1}e_t$.
We believe the beneficiary is two-folded:
First, it should better describe the high-frequency market. De Jong and Schotman's diagonal variance-covariance matrix restricted that all agencies immediately react to the efficient shock. However, various papers have indicated that information arrival can be delayed - some investors react to the market with lag. Various studies of autocorrelation patterns echo this. We are aware that De Jong and Schotman allowed AR(1) in the error term to allow autocorrelation within the asset price change. However, we believe it would be more reasonable that the autocorrelation comes from (1)the lagged efficient price change - some investors react to the news with delay, and (2) past error-correction mechanisms - investors are correcting the past pricing errors for arbitrage opportunities. Mathematically speaking, by allowing the endogenous error correction terms (e.g. Andersen et al. $p_{t-1} - m_{t-1}$), our model captured further autocorrelation pattern and more second moment patterns, e.g. $E[\Delta p_t \Delta p_{t-h}]$. In another word, we discuss the generous case when $e_{i,t}$ for market i potentially contains lagged r_t, lagged $e_{i,t}$,and lagged and current $e_{j,t}$ from market j.
Second, by adding these endogenous error-correction terms, our model solves the identification issue of the De Jong and Schotman's framework - our model no long request the transitory noise term orthogonal to each other. In other words, our model captures additional market patterns without requesting additional constraints.

4 Information Share Limitations

In this section, since both De Jong and Schotman's and our information share models are based on a similar idea - Hasbrouck's information share, we discuss the limitation of this methodology.

Regression Issues Both De Jong and Schotman, 2010 and our model consider the relationship between the innovation in the efficient price and the shocks to individual prices with the following equation:

$$r_t = \gamma_{ols} v_t + \eta_t \tag{67}$$

Naturally, this methodology encounter the issue that this regression might not be valid. By definition, there might be some multicollinearity issue since the v_t are correlated by both models. However, this normally won't be a severe issue since in normal cases, since it is very unlikely that the price innovations of different markets are highly correlated.[14] But still, when the model parameters indicate that the price innovations are highly correlated, we should aware that the information share calculated in this extreme case might not be reliable. In the extreme case, we might need replace this information share methodology with other methodologies.

Information Share Scope In this part, we discuss the trustworthy scope of the information share of De Jong and Schotman and our model – when the information share can be restricted as [0,1].
Both De Jong and Schotman and our information share back to this basic equation(For simplicity, we discuss the case of 2 markets.):

$$\begin{aligned} v_{s,t} &= \gamma_s r_t + \theta e_{c,t} + e_{s,t} \\ v_{f,t} &= \gamma_f r_t + e_{c,t} + e_{f,t} \\ v_{s,t} &= s_t - m_{t-1} \\ v_{f,t} &= f_t - m_{t-1} \end{aligned} \tag{68}$$

Where all of the parameters are scalars.[15]
Or can be written as vectorized representation:

$$\begin{aligned} v_t &= \beta r_t + \Theta e_{c,t} + e_t \\ \beta &= \begin{pmatrix} \gamma_s \\ \gamma_f \end{pmatrix} \\ v_t &= \begin{pmatrix} v_{s,t} \\ v_{f,t} \end{pmatrix} \\ e_t &= \begin{pmatrix} e_{s,t} \\ e_{f,t} \end{pmatrix} \\ \Theta &= \begin{pmatrix} \theta \\ 1 \end{pmatrix} \end{aligned} \tag{69}$$

In both two models, $Cov(r_t, e_{c,t}) = Cov(e_{c,t}, e_{s,t}) = Cov(e_{c,t}, e_{f,t}) = Cov(r_t, e_{s,t}) = Cov(r_t, e_{f,t}) = 0$.
The two cointegrated markets' prices can be described with the equilibrium above, and we know this model's coefficients and variance-covariance structures from the estimation (as discussed in the identification analysis section). The key question would be the proportions of the variance of the efficient price change

[14] If it happens, it usually means that the estimation frequency is not high enough to distinguish the price series shocks.
[15] Our error correction terms will be included in the $e_{c,t}$ terms.

explained by each market.
Then calculate the variance-covariance matrix of v_t as De Jong and Schotman, 2010:

$$\Upsilon = E[v_t, v_t'] = \beta\beta'\sigma_{r_t}^2 + \Theta\Theta'\sigma_{e_{c,t}}^2 + \sigma_{e_t}^2 \tag{70}$$

Then both De Jong and our model discuss the relationship between the innovation of efficient price and shock to individual prices by this linear regression:

$$r_t = \gamma_{ols} v_t + \eta_t \tag{71}$$

As mentioned before, this regression equation determines that the information share estimation methodology is only valid when this regression is valid.
The regression coefficient, $\mathrm{Var}(r_t)$, R^2, and information share follow the same process in De Jong and our model:

$$\begin{aligned} \gamma_{ols}' &= \Upsilon^{-1}\beta\sigma_{r_t}^2 \\ Var(r_t) &= \gamma'\Upsilon\gamma + \sigma_\eta^2 \\ R^2 &= 1 - \sigma_\eta^2/\sigma_{r_t}^2 = \gamma'\Upsilon\gamma + \sigma_\eta^2 = \gamma'\beta \\ IS &= \gamma \circ \beta \end{aligned} \tag{72}$$

Specifically, now we examine when the regression of equation(71) is valid and what is the scope of the information share in the special case of two markets. Specifically, why γ_{ols} is positive in their model and why the information share is scoped as [0,1]. We still from the general model of 2 markets:

$$\begin{aligned} v_{s,t} &= \gamma_s r_t + \theta e_{c,t} + e_{s,t} \\ v_{f,t} &= \gamma_f r_t + e_{c,t} + e_{f,t} \\ v_{s,t} &= s_t - m_{t-1} \\ v_{f,t} &= f_t - m_{t-1} \end{aligned} \tag{73}$$

We write out the Υ^{-1}:

$$\begin{aligned} \Upsilon^{-1} &= (\beta\beta'\sigma_{r_t}^2 + \Theta\sigma_{e_{c,t}}^2\Theta' + \sigma_{e_t}^2)^{-1} \\ &= \begin{pmatrix} \gamma_s^2\sigma_{r_t}^2 + \theta^2\sigma_{e_{c,t}}^2 + \sigma_{e_{s,t}}^2 & \gamma_s\gamma_f\sigma_{r_t}^2 + \theta\sigma_{e_{c,t}}^2 \\ \gamma_s\gamma_f\sigma_{r_t}^2 + \theta\sigma_{e_{c,t}}^2 & \gamma_f^2\sigma_{r_t}^2 + \sigma_{e_{c,t}}^2 + \sigma_{e_{f,t}}^2 \end{pmatrix}^{-1} \\ &= \frac{1}{(\gamma_f^2\sigma_{r_t}^2 + \sigma_{e_{c,t}}^2 + \sigma_{e_{f,t}}^2)(\gamma_s^2\sigma_{r_t}^2 + \theta^2\sigma_{e_{c,t}}^2 + \sigma_{e_{s,t}}^2) - (\gamma_s\gamma_f\sigma_{r_t}^2 + \theta\sigma_{e_{c,t}}^2)(\gamma_s\gamma_f\sigma_{r_t}^2 + \theta\sigma_{e_{c,t}}^2)} \\ &\begin{pmatrix} \gamma_f^2\sigma_{r_t}^2 + \sigma_{e_{c,t}}^2 + \sigma_{e_{f,t}}^2 & -\gamma_s\gamma_f\sigma_{r_t}^2 - \theta\sigma_{e_{c,t}}^2 \\ -\gamma_s\gamma_f\sigma_{r_t}^2 - \theta\sigma_{e_{c,t}}^2 & \gamma_s^2\sigma_{r_t}^2 + \theta^2\sigma_{e_{c,t}}^2 + \sigma_{e_{s,t}}^2 \end{pmatrix} \end{aligned} \tag{74}$$

Notice, the fraction part is always positive because its denominator is $Var(v_{s,t})Var(v_{f,t}) - Cov(v_{s,t})Cov(v_{f,t})$.
Then from the equation above, we can get (the cross terms are canceled here):

$$
\begin{aligned}
\gamma_{ols} &= \Upsilon^{-1}\beta\sigma^2_{r_t} \\
&= \frac{1}{(\gamma_f^2\sigma^2_{r_t}+\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}})(\gamma_s^2\sigma^2_{r_t}+\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}})-(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})} \\
&\begin{pmatrix} \gamma_f^2\sigma^2_{r_t}+\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}} & -\gamma_s\gamma_f\sigma^2_{r_t}-\theta\sigma^2_{e_{c,t}} \\ -\gamma_s\gamma_f\sigma^2_{r_t}-\theta\sigma^2_{e_{c,t}} & \gamma_s^2\sigma^2_{r_t}+\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}} \end{pmatrix}\begin{pmatrix}\gamma_s \\ \gamma_f\end{pmatrix}\sigma^2_{r_t} \\
&= \frac{\sigma^2_{r_t}}{(\gamma_f^2\sigma^2_{r_t}+\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}})(\gamma_s^2\sigma^2_{r_t}+\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}})-(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})} \\
&\begin{pmatrix} \gamma_s(\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}})-\theta\gamma_f\sigma^2_{e_{c,t}} \\ \gamma_f(\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}})-\theta\gamma_s\sigma^2_{e_{c,t}}\end{pmatrix} \\
IS &= \gamma_{ols}\circ\beta \\
&= \frac{\sigma^2_{r_t}}{(\gamma_f^2\sigma^2_{r_t}+\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}})(\gamma_s^2\sigma^2_{r_t}+\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}})-(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})(\gamma_s\gamma_f\sigma^2_{r_t}+\theta\sigma^2_{e_{c,t}})} \\
&\begin{pmatrix} \gamma_s^2(\sigma^2_{e_{c,t}}+\sigma^2_{e_{f,t}})-\theta\gamma_s\gamma_f\sigma^2_{e_{c,t}} \\ \gamma_f^2(\theta^2\sigma^2_{e_{c,t}}+\sigma^2_{e_{s,t}})-\theta\gamma_s\gamma_f\sigma^2_{e_{c,t}}\end{pmatrix}
\end{aligned}
\tag{75}
$$

With the equation above, we can see when and why in some special case that the correlation in our generalized model might be large and the information share might getting negative - for example, when $\sigma^2_{e_{c,t}}$ is significantly larger than $\sigma^2_{e_{s,t}}$ and $\sigma^2_{e_{f,t}}$. In this case, the smaller terms of γ_s and γ_f lead the mapped part of the matrix to negative.

On the other hand, in De Jong and Schotman's case, the information share as follows:

$$
\begin{aligned}
IS &= \frac{\sigma^2_{r_t}}{(\gamma_f^2\sigma^2_{r_t}+\sigma^2_{e_{f,t}})(\gamma_s^2\sigma^2_{r_t}+\sigma^2_{e_{s,t}})-\gamma_s^2\gamma_f^2\sigma^4_{r_t}}\begin{pmatrix}\gamma_s^2\sigma^2_{e_{f,t}} \\ \gamma_f^2\sigma^2_{e_{s,t}}\end{pmatrix} \\
&= \frac{\sigma^2_{r_t}}{\gamma_f^2\sigma^2_{r_t}\sigma^2_{e_{s,t}}+\gamma_s^2\sigma^2_{r_t}\sigma^2_{e_{s,t}}+\sigma^2_{e_{f,t}}\sigma^2_{e_{s,t}}}\begin{pmatrix}\gamma_s^2\sigma^2_{e_{f,t}} \\ \gamma_f^2\sigma^2_{e_{s,t}}\end{pmatrix} \\
&= \frac{1}{\gamma_f^2\sigma^2_{e_{s,t}}+\gamma_s^2\sigma^2_{e_{s,t}}+\frac{\sigma^2_{e_{f,t}}\sigma^2_{e_{s,t}}}{\sigma^2_{r_t}}}\begin{pmatrix}\gamma_s^2\sigma^2_{e_{f,t}} \\ \gamma_f^2\sigma^2_{e_{s,t}}\end{pmatrix}
\end{aligned}
\tag{76}
$$

Easily, we can see why the information share by De Jong and Schotman is positive and why the sum of them is smaller than 1, and why their framework only allows the price innovations across different markets correlated through the over/under-reaction of efficient price change r_t[16].

In summary, the information share in De Jong and Schotman and our framework is strictly scoped in [0,1] when in the equilibrium (equation(68)) restrict the $\theta_{e,c} = 0$. In other words, to limit the information share restricted in [0,1], the

[16] Only in this case they can guarantee the information share in their framework in [0,1]

assumption that the two markets are only correlated through different scales of over/under-reaction to the efficient price r_t is needed. In our generalized model, we allowed the two markets to correlate through additional mechanisms. Thus, we no longer have the mathematical restriction of the information shares strictly in the scope of [0,1]. However, for the general market case (especially the frequency is reasonable, e.g., frequency of 1s for the co-movement of SPY-EMINI markets for most cases), our generalized information share is normally good enough to capture the information share of the two markets. This is because in normal time, the over/under-reaction to the efficient price change is not very large across these two markets, and the co-movement of these two markets is already well captured by the two error-correction mechanisms embedded in our model. For the extreme cases, our quasi Bayesian approach would restrict information share within the scope of [0,1], so that the estimation is reasonable.