



**HAL**  
open science

## When is cross impact relevant?

Victor Le Coz, Iacopo Mastromatteo, Damien Challet, Michael Benzaquen

► **To cite this version:**

Victor Le Coz, Iacopo Mastromatteo, Damien Challet, Michael Benzaquen. When is cross impact relevant?. *Quantitative Finance*, 2024, 24 (2), pp.265. hal-04234766

**HAL Id: hal-04234766**

**<https://hal.science/hal-04234766v1>**

Submitted on 10 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# When is cross impact relevant?

Victor Le Coz,<sup>1,2,3,\*</sup> Iacopo Mastromatteo,<sup>4,†</sup> Damien Challet,<sup>5,‡</sup> and Michael Benzaquen<sup>1,2,4,§</sup>

<sup>1</sup>*Chair of Econophysics and Complex Systems,  
École polytechnique,  
91128 Palaiseau Cedex,  
France*

<sup>2</sup>*LadHyX UMR CNRS 7646,  
École polytechnique,  
91128 Palaiseau Cedex,  
France*

<sup>3</sup>*Quant AI lab, 29 Rue de Choiseul, 75002 Paris,  
France*

<sup>4</sup>*Capital Fund Management,  
23 Rue de l'Université, 75007 Paris,  
France*

<sup>5</sup>*Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes,  
CentraleSupélec, Université Paris-Saclay,  
91192 Gif-sur-Yvette Cedex,  
France*

(Dated: May 29, 2023)

Trading pressure from one asset can move the price of another, a phenomenon referred to as *cross impact*. Using tick-by-tick data spanning 5 years for 500 assets listed in the United States, we identify the features that make cross-impact relevant to explain the variance of price returns. We show that price formation occurs endogenously within highly liquid assets. Then, trades in these assets influence the prices of less liquid correlated products, with an impact velocity constrained by their minimum trading frequency. We investigate the implications of such a multidimensional price formation mechanism on interest rate markets. We find that the 10-year bond future serves as the primary liquidity reservoir, influencing the prices of cash bonds and futures contracts within the interest rate curve. Such behaviour challenges the validity of the theory in Financial Economics that regards long-term rates as agents anticipations of future short term rates.

## I. INTRODUCTION

According to standard economic theory, the price of an asset should integrate all publicly available information regarding its fundamental value. In practice, price formation occurs through a trading system that mechanically forces information flow into prices via the order flow of market participants. This well-established phenomenon is known as *price impact*.

An early model of price impact was proposed by Kyle (1985), who assumed a linear dependence between absolute price differences and signed traded volumes. Further work established that the price impact of large (split) trades universally follows a square-root law of the traded volume (Loeb, 1983; Plerou *et al.*, 2004; Almgren *et al.*, 2005; Kissell and Malamut, 2005; Moro *et al.*, 2009; Toth *et al.*, 2011; Mastromatteo *et al.*, 2014; Donier and Bonart, 2015; Zarinelli *et al.*, 2015; Kyle and Obizhaeva,

2023; Bacry *et al.*, 2015; Tóth *et al.*, 2017; Bouchaud *et al.*, 2018). Yet, the price impact of a single anonymous market order is a much weaker concave (almost constant) function of its volume when the latter is adequately normalized by the available liquidity in the order book (Bouchaud *et al.*, 2018; Hasbrouck, 1991; Chen *et al.*, 2002; Lillo *et al.*, 2003; Potters and Bouchaud, 2003; Zhou, 2012; Gomber *et al.*, 2015). This behavior is due to the *selective liquidity taking* effect (Bouchaud *et al.*, 2018; Taranto *et al.*, 2014): most of the large market order arrivals happen when there is a large volume available at the opposite-side best quote, specifically trying to avoid moving the mid-price. To overcome this effect, impact is often measured over a coarse-grained time scale  $\tau$ , by aggregating trades into a *signed order flow imbalance*. This method involves calculating the signed sum of the volumes of all trades within a time window of length  $\tau$ , while observing price changes during the same interval. Within this framework, the magnitude of price impact crosses over from a linear to a concave behavior, as the signed order flow increases (Bouchaud *et al.*, 2018; Kempf and Korn, 1999; Plerou *et al.*, 2002; Evans and Lyons, 2002; Chordia and Subrahmanyam, 2004; Gabaix *et al.*, 2006; Hopman, 2007; Patzelt and Bouchaud, 2017,

\* victor.lecoz@quant.global

† iacopo.mastromatteo@cfm.com

‡ damien.challet@centralesupelec.fr

§ michael.benzaquen@polytechnique.edu

2018). In addition, the aggregated price impact is a concave function of the time scale  $\tau$  chosen for the aggregation (Bouchaud *et al.*, 2018). Finally, in order to conciliate the long term positive auto-correlation of trades with the independence of price increments, Bouchaud *et al.* (2004) established that price impact must be *transient*. This assumption means that the magnitude of the price-impact of a trade decreases across time. This hypothesis was corroborated in the following years (Tóth *et al.*, 2017; Hopman, 2007; Bouchaud *et al.*, 2006, 2009; Gatheral, 2010; Gatheral and Schied, 2013; Alfonsi *et al.*, 2016; Gârleanu and Pedersen, 2016; Taranto *et al.*, 2018; Ekren and Muhle-Karbe, 2019).

A more subtle effect is that trading pressure from one asset can move the price of another. This effect, which is referred to as *cross-impact*, was studied initially by Hasbrouck and Seppi (2001) and later in Chordia *et al.* (2001); Evans and Lyons (2001); Harford and Kaul (2005); Pasquariello and Vega (2007); Andrade *et al.* (2008); Tookes (2008); Pasquariello and Vega (2015); Wang and Guhr (2017); Benzaquen *et al.* (2017); Schneider and Lillo (2019); Tomas *et al.* (2021, 2022); Brigo *et al.* (2022).

The simplest cross-impact models posit a linear relationship between signed trading volumes and prices variations in time windows of length  $\tau$  (the binning frequency) (Hasbrouck and Seppi, 2001; Harford and Kaul, 2005; Pasquariello and Vega, 2007, 2015; Tomas *et al.*, 2021, 2022). While the time decay of the transient impact model was studied for bonds (Schneider and Lillo, 2019; Schneider, 2019) and stocks (Wang, 2017), the time scale maximizing the accuracy of linear cross-impact models has not yet been documented. Moreover, this optimal time scale is an indicator of the speed of information transmission among assets, which has not been studied extensively, although Zumbach and Lynch (2001); Lynch and Zumbach (2003); Cordi *et al.* (2021) inferred typical time scales of market reactions of the volatility process. In addition, Tomas *et al.* (2022); Rosenbaum and Tomas (2022); Cordoni *et al.* (2022) link the magnitude of cross-impact to asset liquidity and to the correlation among assets.

Here, we quantitatively characterize the circumstances under which a model with cross-impact over-performs one that does not include impact across assets. Additionally, we identify the time scales that maximize the accuracy of linear cross-impact models. Our study includes an introduction to the linear cross-impact modeling framework and a methodology to evaluate the factors influencing cross-impact's relevance in explaining price return variance. The results are organized according to the studied features: the bin size, the trading frequency, the correlation among assets, and the liquidity. In the final section, we provide applications of these findings to the interest rate curve.

## II. NOTATIONS

The set of real-valued square matrices of dimension  $n$  is denoted by  $\mathcal{M}_n(\mathbb{R})$ , the set of orthogonal matrices by  $\mathcal{O}_n$ , the set of real symmetric positive semi-definite matrices by  $\mathcal{S}_n^+(\mathbb{R})$ , and the set of real symmetric positive definite matrices by  $\mathcal{S}_n^{++}(\mathbb{R})$ . Given a matrix  $A$  in  $\mathcal{M}_n(\mathbb{R})$ ,  $A^\top$  denotes its transpose. Given  $A$  in  $\mathcal{S}_n^+(\mathbb{R})$ , we write  $A^{1/2}$  for a matrix such that  $A^{1/2}(A^{1/2})^\top = A$ , and  $\sqrt{A}$  for the matrix square root: the unique positive semi-definite symmetric matrix such that  $(\sqrt{A})^2 = A$ . We also write  $\text{diag}(A)$  for the vector in  $\mathbb{R}^n$  formed by the diagonal items of  $A$ . Finally, given a vector  $v$  in  $\mathbb{R}^n$ , we denote the components of  $v$  by  $(v_1, \dots, v_n)$ , and the diagonal matrix whose components are the components of  $v$  by  $\text{diag}(v)$ . See also the table of notations in Appendix A.

## III. MODELING FRAMEWORK

To relate trades to prices, we observe the mid-prices and market orders of  $n$  different assets, both binned at a regular time interval of length  $\tau$  seconds. We denote by  $p_{t,i}$  the opening price of asset  $i$  in the time window  $[t, t + \tau]$  and by  $p_t = (p_{t,1}, \dots, p_{t,n})$  the vector of asset prices at opening. We define  $q_{t,i}$  as the net market order flow traded during the time window  $[t, t + \tau]$ . This is calculated by taking the sum of the volumes of all trades during that time period, with buy trades counted as positive and sell trades counted as negative. Hence,  $q_t = (q_{t,1}, \dots, q_{t,n})$  is the vector of the net traded order flows.

Following the approach proposed by Tomas *et al.* (2022), we study the relationship between the time series of net order flows  $\{q_0, q_\tau, \dots\}$  and the time series of prices  $\{p_0, p_\tau, \dots\}$ , under the two following assumptions:

- prices changes  $\Delta p_t := p_{t+\tau} - p_t$  and order flow imbalances  $q_t$  are linearly related, i.e.,

$$\Delta p_t = \Lambda_t q_t + \eta_t, \quad (1)$$

where the  $n \times n$  matrix  $\Lambda_t$  is the cross-impact matrix and  $\eta_t = (\eta_{t,1}, \dots, \eta_{t,n})$  is a vector of zero-mean random variables representing exogenous noise;

- the cross-impact matrix  $\Lambda_t$  is a function of the form:

$$\Lambda_t = \Lambda_t(\Sigma_t, \Omega_t, R_t), \quad (2)$$

where  $\Lambda_t : \mathcal{S}_n^+(\mathbb{R}) \times \mathcal{S}_n^+(\mathbb{R}) \times \mathcal{M}_n(\mathbb{R}) \rightarrow \mathcal{M}_n(\mathbb{R})$  is called a cross-impact model,  $\Sigma_t := \text{cov}(\Delta p_t)$  is the price change covariance matrix,  $\Omega_t := \text{cov}(q_t)$  is the order flow covariance matrix, and  $R_t = \mathbb{E}[(\Delta p_t - \mathbb{E}[\Delta p_t])(q_t - \mathbb{E}[q_t])^\top]$  is a response matrix.

We also define the price variation volatility by

$$\sigma_t := (\sqrt{\Sigma_{t,11}}, \dots, \sqrt{\Sigma_{t,nn}}), \quad (3)$$

and the signed order flow volatility by

$$\omega_t := (\sqrt{\Omega_{t,11}}, \dots, \sqrt{\Omega_{t,nn}}). \quad (4)$$

Finally, for a given asset  $i$ , we define the the average across time of its price variation volatility by

$$\bar{\sigma}_i := \langle \sigma_{t,i} \rangle, \quad (5)$$

and the average across time of its signed order flow volatility by

$$\bar{\omega}_i := \langle \omega_{t,i} \rangle. \quad (6)$$

### A. Definition of the models

Let  $Y$  denote a scalar called the *Y-ratio*. We study the following three cross-impact models:

- the diagonal model, defined by

$$\Lambda_{\text{diag}}(\Sigma, \Omega, R) := Y \text{diag}(R) \text{diag}(\Omega^{-1}), \quad (7)$$

which is the limit case where the cross-sectional impact is set to zero;

- the Maximum Likelihood model (*ML model* in the following sections), defined by

$$\Lambda_{\text{ML}}(\Sigma, \Omega, R) := Y R \Omega^{-1}; \quad (8)$$

- and the Kyle model, defined by

$$\Lambda_{\text{Kyle}}(\Sigma, \Omega, R) := Y (\Omega^{-1/2})^\top \sqrt{(\Omega^{1/2})^\top \Sigma \Omega^{1/2}} \Omega^{-1/2}. \quad (9)$$

The Y-ratio is a re-scaling adjustment parameter, estimated by minimizing, across all assets, the squared errors between the price variations predicted by the model and the realized prices.

It is important to note that the diagonal model can be defined by

$$\Lambda_{\text{diag}} := \text{diag}(\lambda_{\text{diag}}), \quad (10)$$

where the vector  $\lambda_{\text{diag}} = (\lambda_1, \dots, \lambda_n)$  is defined by a set of linear equations:

$$\forall i \in \llbracket 1, n \rrbracket, \Delta p_{t,i} = \lambda_i q_{t,i} + \eta_{t,i}. \quad (11)$$

This means that the diagonal model assumes that each asset  $i$  has its own unique relationship between price increments and order flows, as captured by the coefficient  $\lambda_i$ .

The comparison between the last two models and the first one will allow us to distinguish among the portion of cross-impact that is explained by order flow commonality (which the diagonal model can capture) and the contributions that cannot be explained by this effect, thus requiring models such as ML and Kyle.

### B. Properties of the models

As demonstrated by Tomas *et al.* (2022), the previously defined models satisfy a list of properties that characterize their behavior. These properties are recalled below.

1. *Symmetry* properties aim at ensuring that the cross-impact model behaves in a controlled manner under financially-grounded transformations of its variables  $\Sigma_t$ ,  $\Omega_t$ ,  $R_t$ . The Kyle and ML models both adapt to (i) a re-ordering of the considered assets (*permutation invariance*), (ii) a change of currency (*cash invariance*) or (iii) volume units (*split invariance*) and, (iv) a change of basis in the asset space (*rotational invariance*). In contrast, the diagonal model crucially misses property (iv), as it regards the physical space of assets as a privileged basis for the description.
2. *Non-arbitrage* properties aim at ensuring the absence of arbitrage in the sense of (Gatheral, 2010), i.e., round-trip trading strategies with positive average profit. Both the diagonal and Kyle model prevent (i) static arbitrage over a single-period (thanks to their *positive semi-definiteness*) and (ii) dynamic arbitrage over multi-period. Yet, the ML model does not satisfy any of these non-arbitrage properties.
3. *Fragmentation* properties aim at ensuring the equality of the price impacts generated from traded volumes of the same assets on fragmented markets (e.g. US stocks are traded on several venues). This property is satisfied by both the Kyle and ML models but is trivially violated by the diagonal model.
4. *Stability* properties aim at ensuring the impossibility to manipulate the price of liquid products using illiquid instruments. This property is satisfied by the Kyle model and the diagonal model, but not by the ML model.

## IV. METHODOLOGY

### A. Estimation method

We use tick-by-tick trades and quotes for 500 assets quoted in limit order books in the United States. Our sample includes stocks, bonds, futures on bonds and futures on stock indexes. Unless otherwise specified, our data set covers the 2017 – 2022 period. For a given year, we consider the data from the preceding year as in-sample data, while the data from the current year is designated as out-of-sample data. We then aggregate in and out-of-sample results over all years.

A significant portion of our analysis involves the selection of pairs of assets from the pool of the 500 assets in our sample. For this purpose, we select 20,000 pairs of assets per year, and aggregate our results over a five-year period, resulting in 100,000 year-pair couples.

To overcome the conditional heteroskedasticity of price variations and signed order flows, we use a daily estimator of their volatility. Let  $\{t_1, \dots, t_K\} \in \mathbb{R}^K$  denote the  $K$  business days of a year. For each day  $t_k$ , the estimators of the price increments volatility and of the signed order flows volatility are defined by

$$\begin{aligned}\widehat{\sigma}_{t_k} &:= (\langle \Delta p_{t_k,1}^2 \rangle_{t_k}, \dots, \langle \Delta p_{t_k,n}^2 \rangle_{t_k}), \\ \widehat{\omega}_{t_k} &:= (\langle q_{t_k,1}^2 \rangle_{t_k}, \dots, \langle q_{t_k,n}^2 \rangle_{t_k}),\end{aligned}\quad (12)$$

respectively, where the average  $\langle \cdot \rangle_{t_k}$  is computed using data on the day  $t_k$ . We assume the correlation matrices  $\text{diag}(\sigma_t)^{-1} \Sigma_t \text{diag}(\sigma_t)^{-1}$  and  $\text{diag}(\omega_t)^{-1} \Omega_t \text{diag}(\omega_t)^{-1}$ , as well as the normalized response matrix  $\text{diag}(\sigma_t)^{-1} R_t \text{diag}(\omega_t)^{-1}$ , are stationary. Let  $\widehat{\rho}_{\Delta p}$ ,  $\widehat{\rho}_q$  and  $\widehat{\rho}_{\Delta p, q}$  denote their respective estimators using one year of data. Thus, on the day  $t_k$ , the estimated covariance and response matrices  $\widehat{\Sigma}_{t_k}$ ,  $\widehat{\Omega}_{t_k}$  and  $\widehat{R}_{t_k}$  are obtained by

$$\begin{aligned}\widehat{\Sigma}_{t_k} &= \text{diag}(\widehat{\sigma}_{t_k}) \widehat{\rho}_{\Delta p} \text{diag}(\widehat{\sigma}_{t_k}), \\ \widehat{\Omega}_{t_k} &= \text{diag}(\widehat{\omega}_{t_k}) \widehat{\rho}_q \text{diag}(\widehat{\omega}_{t_k}), \\ \widehat{R}_{t_k} &= \text{diag}(\widehat{\sigma}_{t_k}) \widehat{\rho}_{\Delta p, q} \text{diag}(\widehat{\omega}_{t_k}),\end{aligned}\quad (13)$$

respectively.

## B. Metrics definition

### 1. Goodness-of-fit

For a given cross-impact model  $\Lambda_t$ , the predicted price change for the time window  $[t, t + \tau]$  due to the order flow imbalance  $q_t$  is defined as

$$\widehat{\Delta p}_t := \Lambda_t(\widehat{\Sigma}_t, \widehat{\Omega}_t, \widehat{R}_t) q_t. \quad (14)$$

To evaluate the goodness-of-fit of the cross-impact model  $\Lambda$ , we compare the predicted price changes  $\widehat{\Delta p}_t$  to the realized price changes  $\Delta p_t$ . For this purpose, we use a performance indicator parameterized by a symmetric, positive definite matrix  $M \in \mathcal{S}_n^+(\mathbb{R})$ ,  $M \neq 0$ . Let  $\{t_1, \dots, t_N\} \in \mathbb{R}^N$  denote  $N$  sample times,  $\{\Delta p_{t_1}, \dots, \Delta p_{t_N}\}$  be a realization of the price process, and  $\{\widehat{\Delta p}_{t_1}, \dots, \widehat{\Delta p}_{t_N}\}$  denote the corresponding series of predictions from the model. The  $M$ -weighted generalized  $\mathcal{R}^2(M)$  is defined as

$$\mathcal{R}^2(M) := 1 - \frac{\sum_{i=1}^N (\Delta p_{t_i} - \widehat{\Delta p}_{t_i})^\top M (\Delta p_{t_i} - \widehat{\Delta p}_{t_i})}{\sum_{i=1}^N \Delta p_{t_i}^\top M \Delta p_{t_i}}. \quad (15)$$

The closer this score is to one, the better is the fit to the actual prices. To highlight different sources of error, different choices of  $M$  can be considered:

- $M = I_\sigma := \text{diag}(\langle \sigma_t^2 \rangle)^{-1}$ , to account for errors relative to the typical deviation of the asset considered. This type of error is relevant for strategies predicting idiosyncratic moves of the constituents of the basket, rather than strategies betting on correlated market moves.
- $M = I_{\sigma_i} := \text{diag}(\langle \sigma_{t,i}^2 \rangle)^{-1}$ , to account for errors of a single stock  $i$ .
- $M = \langle \Sigma_t \rangle^{-1}$ , to consider how well the model predicts the individual modes of the return covariance matrix. This would be the relevant error measure for strategies that place a constant amount of risk on the modes of the correlation matrix, leveraging up combinations of products with low volatility and scaling down market direction that exhibit large fluctuations.

Within the following sections, we mainly study the cross-impact goodness-of-fit for pairs  $(i, j)$  of assets. In these cases, unless stated otherwise, we calculate  $\mathcal{R}^2(I_{\sigma_i})$  to measure solely the errors on the first asset  $i$ , the predicted asset, as a function of the characteristics of the second asset  $j$ , the explanatory asset.

Additionally, we define a second indicator to determine the extent to which the goodness-of-fit results from cross-sectional information. We define  $\Delta \mathcal{R}^2(M)$ , the accuracy increase from the cross sectional model as

$$\Delta \mathcal{R}^2(M) := \mathcal{R}^2(M) - \mathcal{R}_{\text{diag}}^2(M), \quad (16)$$

where  $\mathcal{R}_{\text{diag}}^2(M)$  is the  $M$ -weighted generalized  $\mathcal{R}^2(M)$  obtained from the degenerated model without cross-sectional impact  $\Lambda_{\text{diag}}$ .

Tests that confirm the statistical significance of the  $\mathcal{R}^2$  in the case of a single asset are reported in the appendix B.1. Nevertheless, appendix B.2 demonstrates that the auto-correlation of signed order flows invalidates the linear framework used in this study. Using of a more accurate *propagator model* (Bouchaud *et al.*, 2018, 2006; Alfonsi *et al.*, 2016; Benzaquen *et al.*, 2017; Schneider and Lillo, 2019; Bouchaud, 2009) would yield only marginal improvements in the goodness-of-fit (Tomas *et al.*, 2022) but it would impede conducting this study at the same scale across time and assets.

### 2. Definition of the assets characteristics

Several assets characteristics are investigated in this study:

- the *bin size*  $\tau$ ;

- the *trading frequency*  $f$ , defined by the number of trades per second;
- the *price increments correlation*  $\rho_{ij}$  between the assets  $i$  and  $j$ ;
- the *liquidity*, defined by the risk of profit or loss in monetary units  $\bar{\omega}_i \bar{\sigma}_i$  over a given time window.

These metrics are further described at the beginning of each corresponding sub-sections below.

## V. RESULTS

### A. The effect of the bin size

Within the linear framework previously defined, the market price impact of a single asset is a non-trivial function of the bin size  $\tau$  (Fig. 1). Specifically, the goodness-of-fit  $\mathcal{R}^2(I_{\sigma_i})$  increases with the bin size up to a maximum ranging generally between 10 and 100 seconds before decreasing down to a negligible level at 1 hour.

At short time scales, effects similar to the Epps effect (Epps, 1979; Toth and Kertesz, 2009) may prevent the correlation between the order flow and the price impact to become fully apparent without further corrections. Indeed, the correlation between the signed order flows and the price variations  $\frac{R_{ii}}{\sigma_i \bar{\omega}_i}$  decreases when the time scale shortens (Fig. 1). In fact, this correlation is simply the square-root of the model accuracy  $\mathcal{R}^2(I_{\sigma_i})$ . Yet, among the causes of the Epps effect (lead-lag effects, asynchronicity, and the minimal response time of traders) (Toth and Kertesz, 2009), only the third factor is deemed relevant. Indeed, for a given asset, its signed order flows and prices are updated synchronously with no lead-lag. More prosaically, when the bin size widens, the number of trades per bin increases and so does the accuracy of the model.

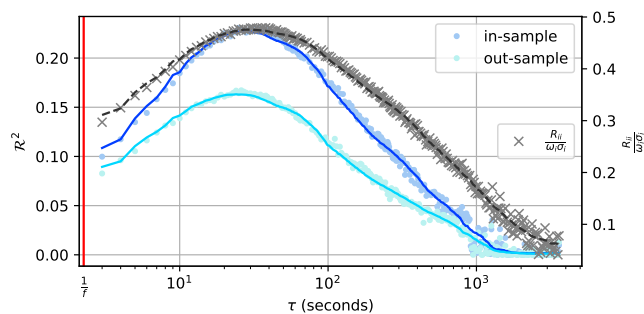


FIG. 1: Single asset generalised R-squared  $\mathcal{R}^2(I_{\sigma_i})$  as a function of the bin size  $\tau$  for the asset *TRMB* (*Trimble Navigation*). In-sample results were calibrated on the year 2021, while out-of-sample outcomes cover 2022. The vertical red line indicates the average time interval between two trades  $1/f$ .

At larger time scales, the predictive power of the linear impact model decreases rapidly (Fig. 1). This decay cannot be attributed to the transient effect from the price impact (see section I). Indeed, the magnitude of the price impact follows a power law with a slow decline, typically remaining significant after a thousand trades (Bouchaud *et al.*, 2018). In the above example, the number of trades accumulated within one hour (the largest bin) is around 1400. Consequently, it is reasonable to anticipate that the impact of the first trades within a bin would continue to be substantial at this time scale. However, we observe a decrease in the price impact model accuracy after a couple of minutes.

More generally, the impact from all the other trades, including the most recent ones, should be observable on the current price change in a bin, even at large time scales. However, as documented by Patzelt and Bouchaud (2017, 2018), the relationship between the signed order flows and the price changes is actually reasonably fitted by a sigmoid function. This latter is indeed linear for reasonably small sizes of signed order flows. Yet, the non-linear relationship between price impacts and larger sizes of signed order flows, which is frequently observed in large-scale bins, explains the lower precision of linear models at these scales.

The relationship between the bin size  $\tau$  and the model accuracy  $\mathcal{R}^2(M)$  provides an avenue to determine the maximum goodness-of-fit  $\mathcal{R}^{2*}(M)$  and its corresponding optimal time scale  $\tau^*(M)$ . The ensuing sections investigate how these latter are influenced by the trading frequency of the assets, correlation among assets, and liquidity of the assets.

### B. The effect of the trading frequency

#### 1. Time scales

For a given asset, we define its trading frequency  $f$  as its average number of trades per second during open market hours. Intuitively, one could expect higher trading frequencies to be associated with shorter optimal time scales  $\tau^*$ , due to the quickest accumulation of trades in a bin. Empirically, higher trading frequencies do decrease the minimum optimal time scales achievable, yet other factors cause the considered assets to deviate from this limit. Specifically, the envelope of the density plot associating optimal time scales to trading frequencies does not contain the lower left section of Fig. 2. Thus, a minimum number of trades is required to reach the optimal bin size. Specifically, the blue straight line on this figure represents the function  $f \rightarrow 1/f$ . Hence, one can estimate this minimum number of trades as the intercept ensuring that the majority of the data points are above this line. We find that a minimum of 10 to 20 trades is required to reach the optimal time scale of a linear

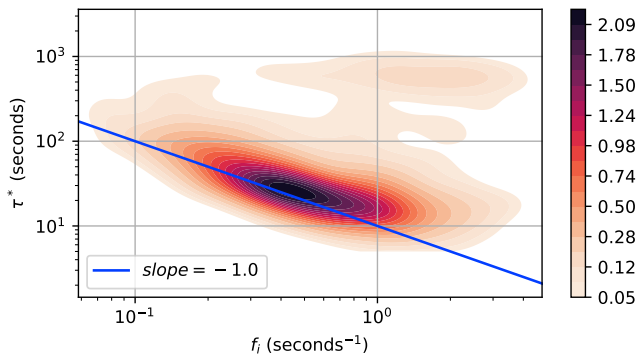


FIG. 2: Empirical distribution of the optimal time scale out-of-sample  $\tau^*(I_{\sigma_i})$  as a function of the trading frequency  $f$  for single assets.

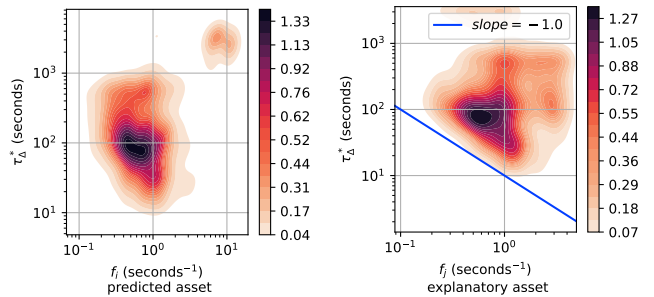
cross-impact model.

The light brown area at the top right of Fig. 2 reveals a smaller group of assets with maximum cross-impact accuracy  $\mathcal{R}^{2*}(I_{\sigma_i})$  for extended time scales. This group includes mainly large capitalization stocks (e.g. AMZN, AAPL, TSLA) and highly traded futures (E-mini S&P futures, 10-year bond futures). For these assets, the causes decreasing the accuracy of the model seems to become effective for a larger number of trades (500 to 5000 trades).

Figure 3b shows the impact of the trading frequency of the explanatory asset  $j$  regarding the goodness-of-fit on asset  $i$ , when calibrating the model on pairs of assets. Specifically,  $\tau_{\Delta}^*(I_{\sigma_i})$  corresponds to the time scale maximizing the added accuracy,  $\Delta\mathcal{R}^{2*}(I_{\sigma_i})$ , when predicting asset  $i$ 's price increments. This indicator is driven by the trading frequency of the explanatory asset, as shown by the triangle form of Fig. 3b. As expected, this behavior cannot be observed when one increases the trading frequency of the predicted asset (Fig. 3a).

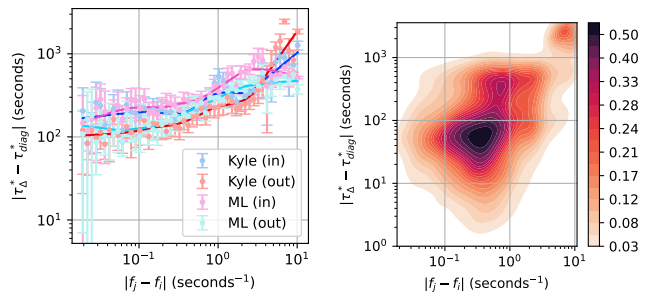
On one hand, the optimal time scale for the cross-sectional effect is influenced by the trading frequency of the explanatory asset. On the other hand, the optimal time scale in the diagonal model is influenced by the frequency of the predicted asset. Consequently, significant deviations between the optimal time scales of the diagonal and cross section models are expected. As depicted in Fig. 4, these deviations increase when the trading frequencies of the two assets diverge. This may decrease the relevance of linear cross-impact models. Indeed, the optimal time scale of the cross sectional impact may be reached when the loss of accuracy from the direct price impact is higher than the marginal gain. However, this effect remains limited as demonstrated in Section VI.B.

Finally, Fig. 5b demonstrates that the time scale  $\tau_{\Delta}^*(I_{\sigma_i})$  maximizing the added accuracy, is affected by the minimum of the trading frequencies of the assets pair. In contrast, the maximum of these two



(a) As a function of  $f_i$ , the trading frequency of the predicted asset. (b) As a function of  $f_j$ , the trading frequency of the explanatory asset.

FIG. 3: Empirical distribution of the optimal time scale out-of-sample  $\tau_{\Delta}^*(I_{\sigma_i})$  as a function of the trading frequency of the predicted asset  $f_i$  (Fig. 3a) or the explanatory asset  $f_j$  (Fig. 3b). The asset pairs are filtered on the 7% of the sample exhibiting a correlation higher than 50% (see section V.C).



(a) Mean deviation  $|\tau_{\Delta}^*(I_{\sigma_i}) - \tau_{\text{diag}}^*(I_{\sigma_i})|$  by buckets of trading frequency gaps  $|f_j - f_i|$ . (b) Empirical distribution of the deviation  $|\tau_{\Delta}^*(I_{\sigma_i}) - \tau_{\text{diag}}^*(I_{\sigma_i})|$  in the Kyle model.

FIG. 4: Asynchronicity of the diagonal and cross sectional models  $|\tau_{\Delta}^*(I_{\sigma_i}) - \tau_{\text{diag}}^*(I_{\sigma_i})|$ , aggregated by buckets of trading frequency gaps (Fig. 4a) or unaggregated (Fig. 4b). The asset pairs are filtered on the 7% of the sample exhibiting a correlation higher than 50% (see section V.C).

frequencies has little effect on this time scale (Fig. 5a). We observe the same behavior with respect to the optimal time scale  $\tau^*(I_{\sigma_i})$  (Fig. 6). Consistently with the single asset case, we find that a minimum of 10 to 20 trades in both assets is required to reach the optimal time scale of a two-dimensional cross-impact model.

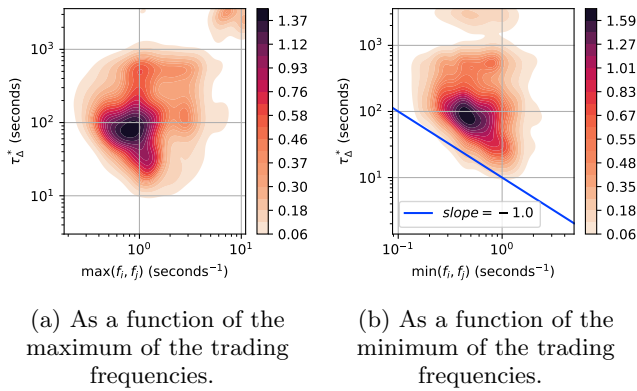


FIG. 5: Empirical distribution of the optimal time scale out-of-sample  $\tau_{\Delta}^*(I_{\sigma_i})$  as a function of the maximum (Fig. 5a) or the minimum (Fig. 5b) of the trading frequencies of the assets pairs. The asset pairs are filtered on the 7% of the sample exhibiting a correlation higher than 50% (see section V.C).

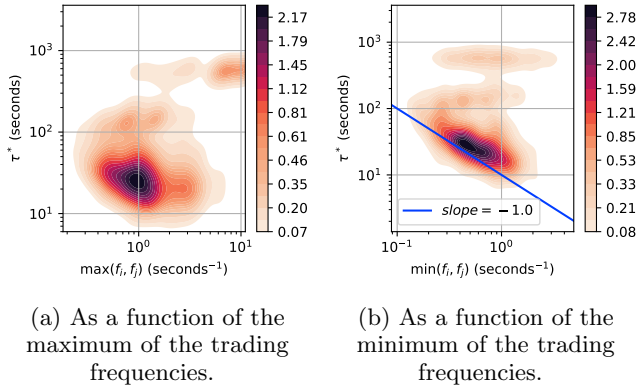


FIG. 6: Empirical distribution of the optimal time scale out-of-sample  $\tau^*(I_{\sigma_i})$  as a function of the maximum (Fig. 6a) or the minimum (Fig. 6b) of the trading frequencies of the assets pairs. The asset pairs are filtered on the 7% of the sample exhibiting a correlation higher than 50% (see section V.C).

## 2. Goodness-of-fit

The trading frequency has a positive effect on the maximal accuracy  $\mathcal{R}^{2*}(I_{\sigma_i})$  observed across the tested bin sizes. Indeed, Fig. 7 shows that an increase in the trading frequency improves the mean  $\mathcal{R}^{2*}(I_{\sigma_i})$  per bucket. Here, the bars shaded in light pastel colors denote the range of two standard deviations surrounding the mean  $\mathcal{R}^{2*}(I_{\sigma_i})$  of the assets bucketed by trading frequency. The continuous lines in bright colors are the *Locally Weighted Scatterplot Smoothing* (Cleveland, 1979) of these mean values. The following figures portraying bucketed data conform to the same convention.

The higher accuracy of the cross-impact model on

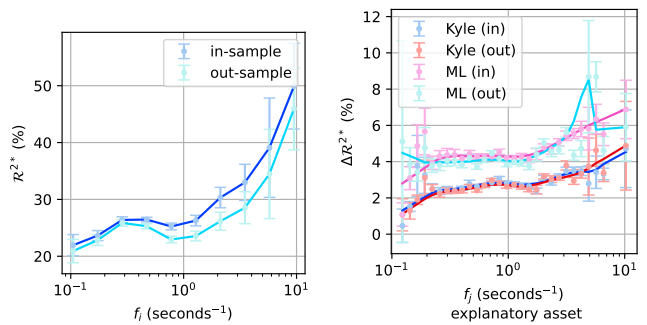


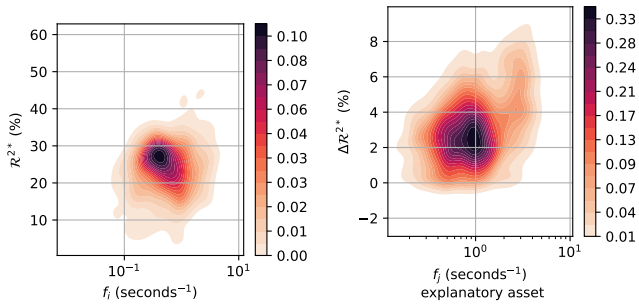
FIG. 7: Mean per trading frequency bucket of the optimal goodness-of-fit  $\mathcal{R}^{2*}(I_{\sigma_i})$  for single assets (Fig. 7a) and pairs of assets (Fig. 7b).

highly traded assets can be attributed to the stronger correlation between prices and order flows when a sufficiently large number of market participants ensure the consistency of the two. However, it must be underlined that the impact of trading frequency is partially offset by observing the optimal accuracy  $\mathcal{R}^{2*}(I_{\sigma_i})$  across bin sizes, resulting in a relatively stable number of trades across trading frequencies. This effect probably explains the low slope in Fig. 7 when excluding data points with large error bars.

Notice also the ML model out-of-sample performance in Fig. 7b is significantly larger than that of the Kyle model, as previously reported in Tomas *et al.* (2022). The ML model can be easily over-fitted if one uses too little data, as it is more flexible than the Kyle model that imposes a no arbitrage condition. The fact that the out-of-sample performance of ML is better than Kyle shows that: (i) frictions in the market (bid-ask spread, fees) at least partially spoil the no-arbitrage assumption, as documented in Schneider (2019); (ii) this effect is significant enough to generalize well to yet unseen data.

To provide a broader view of the distribution of  $\mathcal{R}^{2*}(I_{\sigma_i})$  across trading frequencies in our sample, we also present this data by density in Fig. 8. The figure shows that most assets exhibit a trading frequency around 0.5 trades per second, with an  $\mathcal{R}^{2*}(I_{\sigma_i})$  of 25%.





(a) Single assets in the diagonal model. (b) Pairs of assets in the Kyle model. The asset pairs are filtered on the 7% of the sample exhibiting a correlation higher than 50%.

FIG. 8: Empirical distribution of the optimal out-of-sample  $\mathcal{R}^{2*}(I_{\sigma_i})$  as a function of the trading frequency  $f$  for single assets (Fig. 7a) and pairs of assets (Fig. 8b).

### C. The effect of the correlation among assets

#### 1. Goodness-of-fit

As previously mentioned, correlations among assets are dependent on the time scale at which prices are sampled (Epps, 1979; Toth and Kertesz, 2009; Renò, 2003). Therefore, we use a bin size sufficiently large for the Epps effect to be negligible. Relying on the analysis presented on Fig. 9, we choose a 5-minute bin. This time scale is a good compromise between Epps effect and noise.

As expected, we observe a positive and monotonous relationship between the added accuracy from the cross sectional information  $\Delta\mathcal{R}^{2*}(I_{\sigma})$  and the correlation  $\rho_{ij}$  among pairs of assets. For both the Kyle and ML models,  $\Delta\mathcal{R}^{2*}(I_{\sigma})$  increases from 0 to above 5% over the range of correlation levels in our sample (Fig. 10). Regarding Fig. 10a and the following density plots, the continuous line slope represents the Theil-Sen estimator (Theil, 1992; Sen, 1968; Siegel, 1982), while the dotted lines indicate

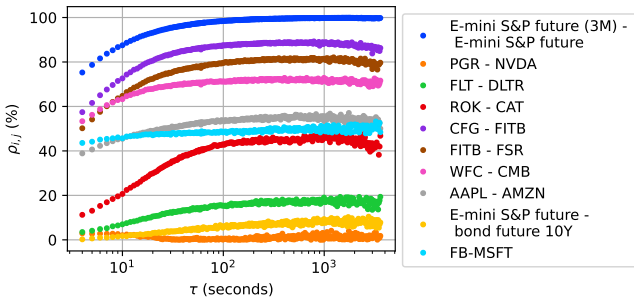
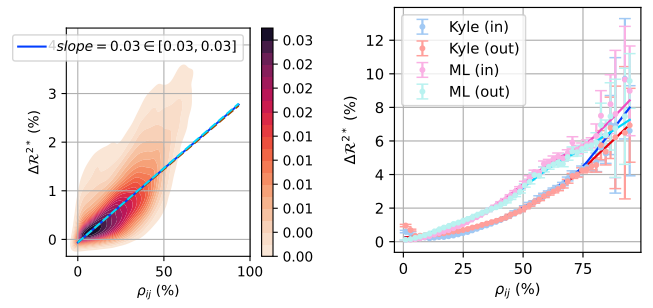
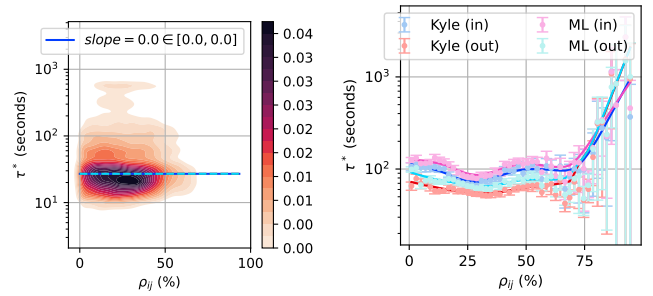


FIG. 9: Linear correlations coefficients  $\rho_{i,j}$  as a function of the bin size  $\tau$ , for a selection of assets couples.



(a) Empirical distribution of the out-of-sample  $\Delta\mathcal{R}^{2*}(I_{\sigma})$  in the Kyle model. (b) Mean per correlation bucket of the optimal  $\Delta\mathcal{R}^{2*}(I_{\sigma})$ .

FIG. 10: Optimal  $\Delta\mathcal{R}^{2*}(I_{\sigma})$ , accounting for errors relative to both assets, as a function of the correlation  $\rho_{ij}$  for pairs of assets.



(a) Empirical distribution in the Kyle model. (b) Mean per correlation bucket.

FIG. 11: Optimal time scale  $\tau^*(I_{\sigma})$  as a function of the correlation level  $\rho_{ij}$  for pairs of assets.

the 95% confidence interval around this estimate.

#### 2. Time scales

The optimal time scale  $\tau^*(I_{\sigma})$  seems unaffected by the correlation level  $\rho_{ij}$  among pairs of assets. Indeed, Fig. 11b shows that the mean value of  $\tau_{\Delta}^*(I_{\sigma})$  is almost independent from the correlation level.

### D. The effect of the liquidity

#### 1. Goodness-of-fit

The liquidity of each individual asset is measured using a risk indicator that represents the typical size of gains or losses during a given time interval. Specifically, the liquidity of the asset  $i$  is defined by  $\bar{\omega}_i \bar{\sigma}_i$ , estimated at a given bin size. We set the bin size to 5 minutes, consistently with the binning frequency for the correlation estimation (see Section V.C).

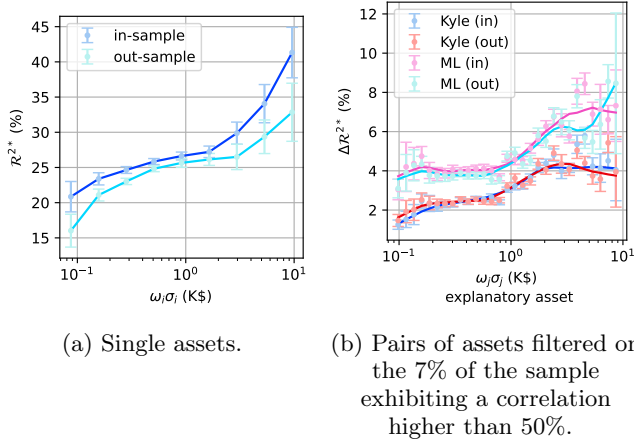


FIG. 12: Mean  $\mathcal{R}^{2*}(I_{\sigma_i})$  by liquidity bucket for single assets (Fig. 12a) and pairs of assets (Fig. 12b).

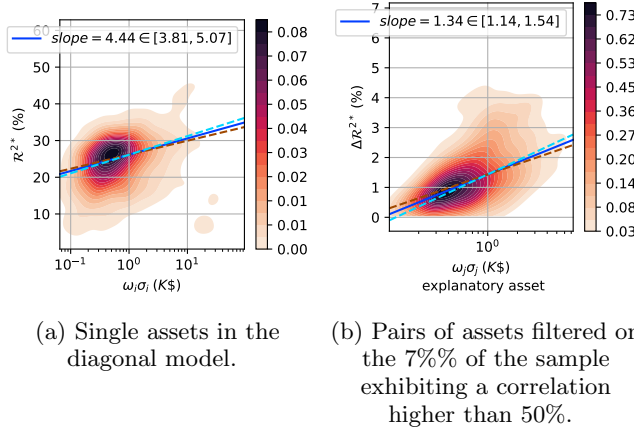


FIG. 13: Empirical distribution of the out-of-sample  $\mathcal{R}^{2*}(I_{\sigma_i})$  as a function of the liquidity, for single assets (13a) and pairs of assets (13b).

Liquidity has a positive effect on the accuracy of the cross-impact models tested, both for single assets and pairs of assets. Indeed, we observe that the out-of-sample goodness-of-fit increases from below 20% to around 30% across the liquidity levels in our sample for the diagonal model (Fig. 12a). Like the interpretation proposed in Section V.B, the higher score on liquid assets can be explained by the stronger correlation between prices and order flows when market liquidity is sufficient to ensure their consistency. As previously, we also present these results through density plots in Fig. 13. These plots illustrate that most assets exhibit a liquidity level of 500 USD per 5 minutes and an  $\mathcal{R}^{2*}(I_{\sigma_i})$  of around 25%.

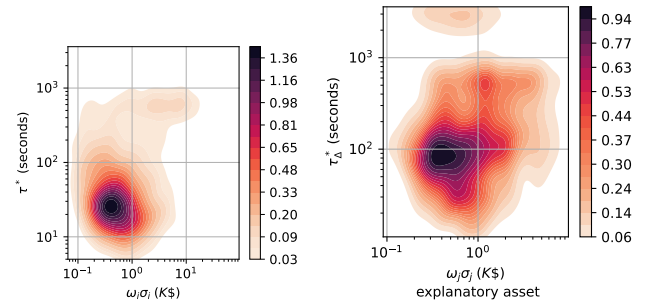


FIG. 14: Empirical distribution of the optimal bin size  $\tau^*(I_{\sigma_i})$  as a function of the liquidity, for single assets (Fig. 14a) and pairs of assets (Fig. 14b).

## 2. Time scales

In contrast, liquidity has an ambiguous effect on the optimal time scale  $\tau^*(I_{\sigma_i})$ . Notably, Fig. 14b exhibits two groups of assets: (i) a large group with medium liquidity and time scales around 90 seconds, (ii) a smaller group with higher liquidity and time scales around 10 minutes. Within both groups, the liquidity seems to have a limited effect on the optimal time scale. These results suggest that other underlying properties of the considered assets influence the optimal time scale.

## 3. Cross effects of individual assets' liquidity

Our objective is to evaluate whether the added accuracy in the Kyle model is consistent with the stability properties outlined in Section III.B. Specifically, we seek to determine whether  $\Delta\mathcal{R}^2(I_{\sigma_i})$  decreases as stock  $i$ 's liquidity increases, but increases as stock  $j$ 's liquidity increases, for a given pair  $(i, j)$  of assets. Notably, we expect that incorporating trades information from a high-liquidity asset to predict the prices of a low-liquidity asset will significantly enhance accuracy. However, this effect is not evident in Fig. 15 due to the cross-effect of the correlation among assets. In fact, the greatest increase in accuracy for predicting asset  $i$ 's prices is observed for relatively high levels of liquidity of asset  $i$ . Empty cells in Fig. 15 and the following heatmaps correspond to either an absence of assets in the associated buckets or to filtered-out results due to measurement errors being greater than 50% of the mean value. As previously, these errors are defined by two standard deviations.

Thus, we neutralize the effect of the correlation in Fig. 16 by grouping the asset pairs of our sample into correlation buckets. While at low correlations levels (from

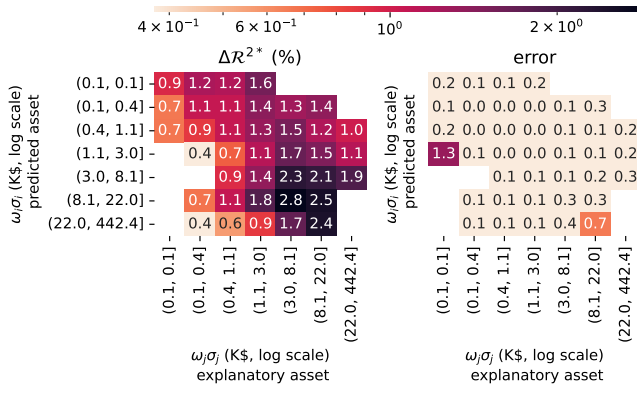


FIG. 15: Mean out-of-sample added accuracy on asset  $i$   $\Delta\mathcal{R}^2(I_{\sigma_i})$ , as a function of the individual risk levels of each asset in the Kyle model.

0% to 20%) the lowest liquid asset remains poorly predicted by the highest liquid asset, the effect of the cross sectional information becomes significant when looking at pairs of well correlated assets (above 50%). In a nutshell, cross-impact is significant if the predicted asset has a lower liquidity than the explanatory asset.

## E. Discussion

Cross-impact is not relevant under all circumstances. Following this analysis, one can establish three requirements to accurately predict prices from cross sectional trades. Firstly, cross-impact does not occur at every time scale. A minimum number of trades in both assets, between 10 to 20, is required to observe a significant added accuracy on the prediction of asset prices. Secondly, trades only significantly explain the prices of highly correlated assets (correlations higher than 50%), regardless of the time scale. Thirdly, cross-impact explains a larger share of price variances if the predicted asset has a lower liquidity than the explanatory asset.

To conclude, we can draw the following narrative. Price formation occurs endogenously within highly liquid assets. Then, trades in these assets influence the prices of their less liquid correlated products, with an impact speed constrained by their minimum trading frequency.

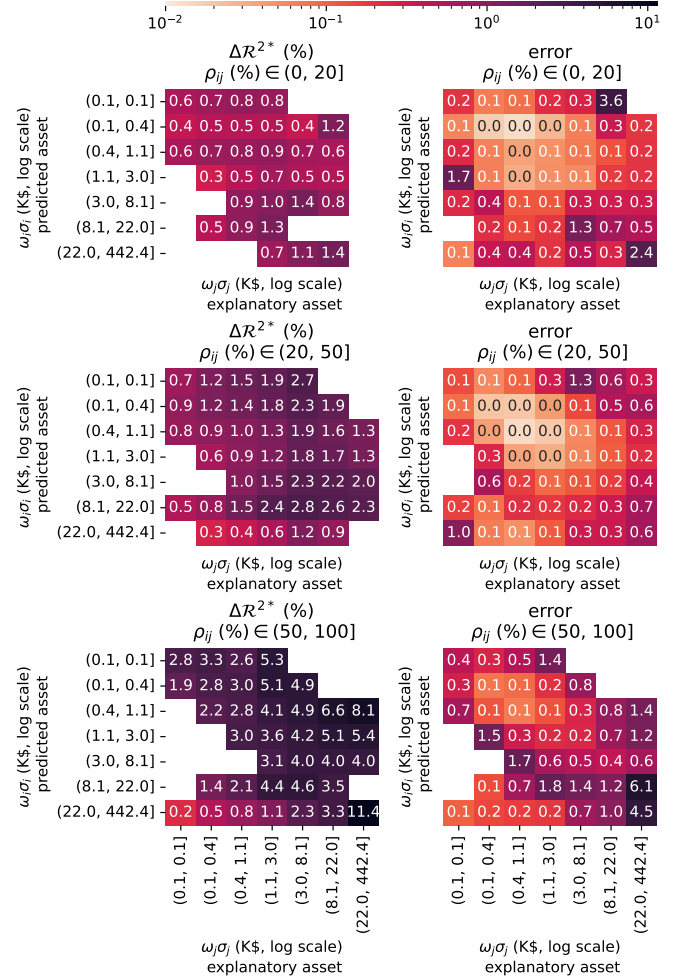


FIG. 16: Mean out-of-sample added accuracy on asset  $i$   $\Delta\mathcal{R}^2(I_{\sigma_i})$ , as a function of the individual risk levels of each asset in the Kyle model.

## VI. APPLICATION TO THE INTEREST RATE CURVE

### A. Assets pairs

According to the previously established narrative, the interest rate curve should be a good candidate to apply a cross-impact model. Indeed, bonds of different tenors are highly correlated and display a wide range of liquidity levels. In this context, we run our previous analysis on a restriction of our initial sample to sovereign cash bonds and bonds futures in the United States for the 2021–2022 period. Figure 17 shows that including the trades of the most liquid assets (the 10-year future and cash bond) significantly increases the prediction accuracy concerning most of the other less liquid tenors.

Of particular significance, Fig. 17 reveals that the trading information transmission flows from the most liquid tenors to that of lower liquid. This behavior challenges the validity of the theory in Financial Economics

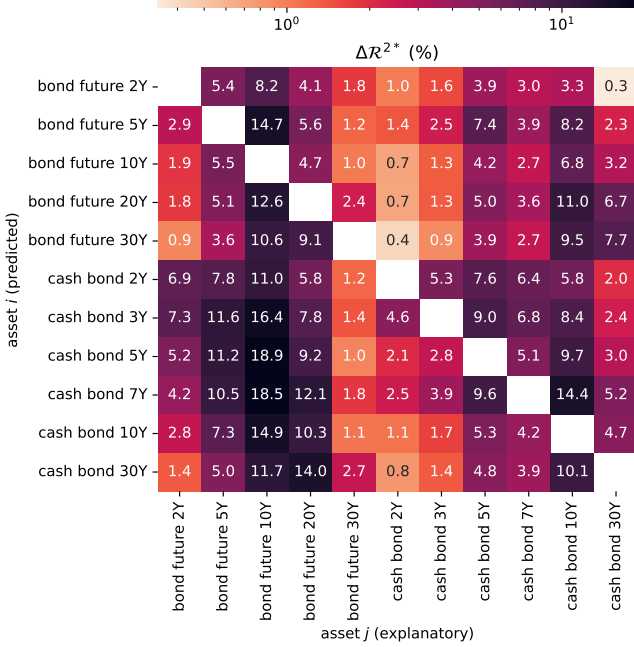


FIG. 17: Out-of-sample added accuracy  $\Delta\mathcal{R}^2(I_{\sigma_i})$  for each pair of assets of the interest rate curve.

that regards long-term rates as agents' anticipations of future short-term rates. In practice, the prices of the low-liquidity tenors are more strongly impacted by the trades of the high-liquidity tenors than vice-versa (e.g. the 2-year cash bond in Fig. 18). Future work could be devoted to the extension of this analysis to repurchase agreements of shorter tenors (such as 1-day, 1-week and 1-month tenors).

Because of the correlation among assets, the total added accuracy from using all asset trades information is not the summation of a given row of the matrix displayed in Fig. 17. Therefore, we display the multidimensional case in the next section.

## B. Multidimensional case

To measure the contribution of each explanatory asset in the multidimensional case, we run the Kyle model using an increasing number of instruments. The results are presented in Fig. 18. Each matrix item  $\mathcal{R}_{ij}^2$  corresponds to the out-of-sample goodness-of-fit  $\mathcal{R}^2(I_{\sigma_i})$  regarding the prediction of asset  $i$  from the set  $\{1, \dots, j\}$  of explanatory assets. In the case of diagonal items, each  $\mathcal{R}_{ii}^2$  represents the effect of the diagonal model using the explanatory asset  $i$ . For example, the row *cash bond 5Y* can be understood as follows: its own trades explain 11.8% of its price increments variance, but the contribution of the other assets increases this score from 14.1% (using the 2-year cash bond) to 45.2% (using all assets).

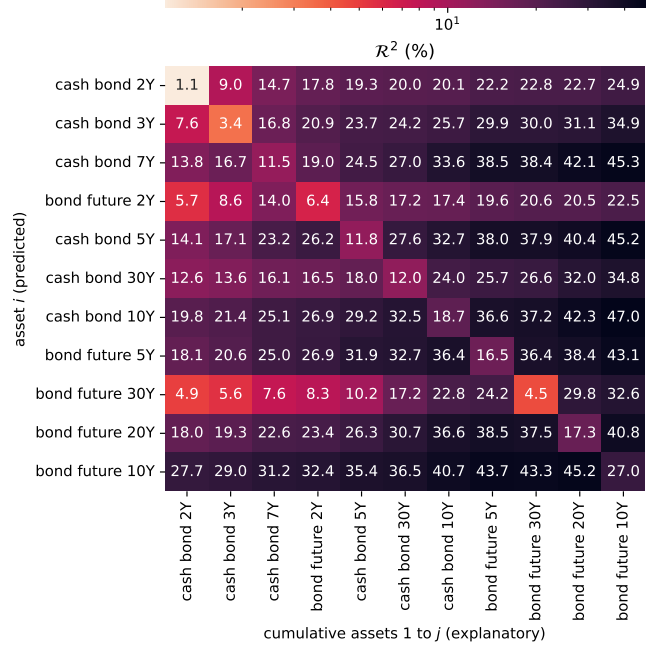


FIG. 18: Out-of-sample goodness-of-fit  $\mathcal{R}^2(I_{\sigma_i})$  for an increasing number of explanatory assets. The bin size was set to 30 minutes, which is close to the optimal time scale for these assets.

Thus, Fig. 18 shows that we can significantly increase the explanatory power of even the most liquid asset when using a sufficiently large number of instruments of lower liquidity.

## C. Kyle matrix analysis

Figure 19 displays the Kyle matrix on 9 November 2021 under two different normalization conventions.

First, Fig. 19a exhibits the Kyle matrix normalized by assets' mean prices  $\frac{\lambda_{i,j}}{\bar{p}_i \bar{p}_j}$ , where  $\bar{p}_j = \langle p_{t,j} \rangle_t$ . Thus, it defines the relative estimated price impact  $\frac{\widehat{\Delta p}_{t,i}}{\bar{p}_i}$  on asset  $i$  from the traded volumes in dollars  $\bar{p}_j q_{t,j}$  on asset  $j$ . Indeed, one can rewrite equation 2 as

$$(\text{diag } \bar{p})^{-1} \widehat{\Delta p}_t = (\text{diag } \bar{p})^{-1} \Lambda_t (\text{diag } \bar{p})^{-1} \text{diag } \bar{p} q_t. \quad (17)$$

However, this re-scaling results in over-weighting the longest tenors. Indeed, for a given interest rate  $r$ , the price of a zero-coupon bond contract of tenor  $T$  and notional  $N$  can be written as  $p = \frac{N}{(1+r)^T} \approx -\frac{N}{rT}$  for  $r \ll 1$  and  $rT \gg 1$ . Consequently, the bond or future contract value decreases linearly with the tenor, so the normalized cross-impact matrix coefficients  $\frac{\lambda_{i,j}}{\bar{p}_i \bar{p}_j}$  are proportional to the squared tenor  $T^2$ . This effect explains the regularities observed in Fig. 19a.

To neutralize the effect of the maturity, we propose a second normalization approach. Fig. 19b presents the

Kyle matrix items normalized by the opposite of the tenor on the left and by the product of the price and tenor on the right:  $-10 \frac{\lambda_{i,j}}{T_i \bar{p}_j T_j}$ , where  $T_i$  is the tenor of the bond  $i$ . These values define the absolute change in the interest rate  $-\frac{\Delta p_{t,i}}{T_i}$  of the bond  $i$  from the traded volumes (in USD) in equivalent 10-year contract in the bond  $j$ :  $\frac{T_j}{10} \bar{p}_j q_{t,j}$ . Formally, it is the reformulation of equation 2 as

$$\begin{aligned}
 & -(\text{diag}(T))^{-1} \widehat{\Delta p}_t = \\
 & -(\text{diag}(T))^{-1} \Lambda_t (\text{diag}(T) \text{diag}(\bar{p}))^{-1} \text{diag}(T) \text{diag}(\bar{p}) q_t.
 \end{aligned}
 \tag{18}$$

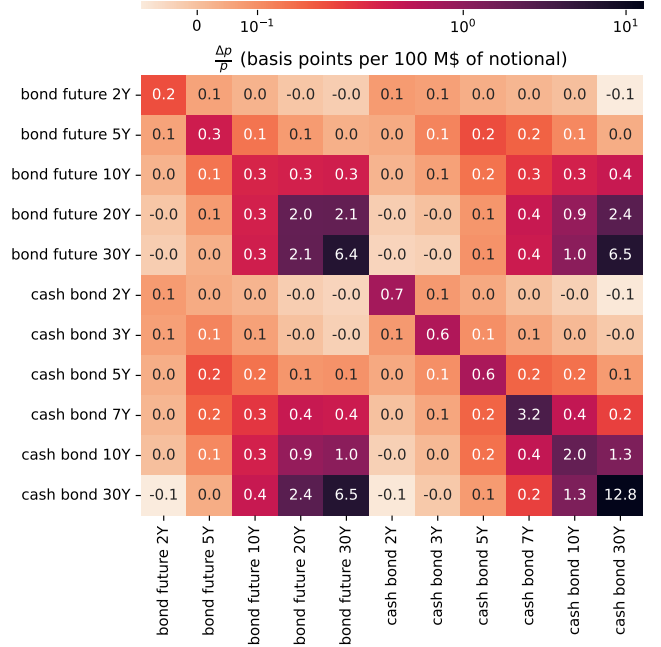
This second approach neutralizes the effect of the tenor on both input volumes and observed prices. Thus, the remaining differences among assets are notably due to correlation and liquidity levels.

In this context, we are able to make four observations from Fig. 19b.

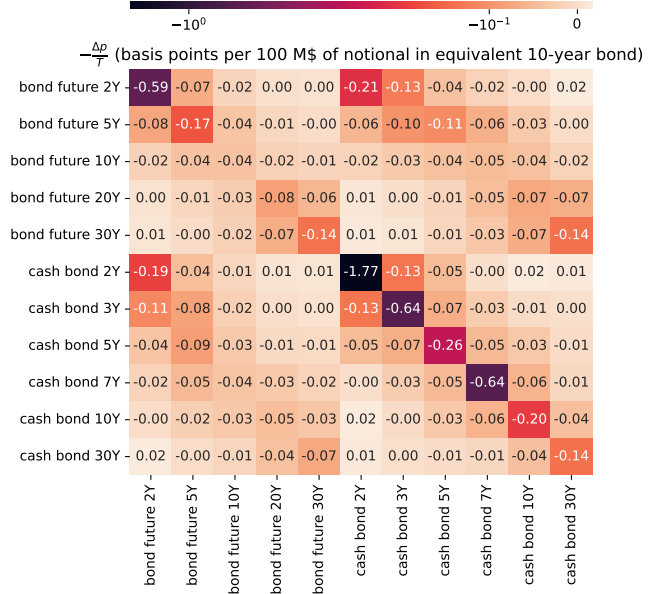
1. Overall impacts are of similar sign and magnitude across all assets, which highlights the first factor of the interest rate curve, the *parallel shift* (Brigo and Mercurio, 2006), due to high correlations.
2. Volumes traded on futures of a given tenor affect more significantly the interest rates of the closest tenors, which shows that correlations are higher among assets of close maturity. Equivalently, it exhibits the structure of other factor(s) beyond the parallel shift.
3. We observe a similar behavior for cash and futures contracts, because of the high correlations between an underlying and its derivative.
4. As a last observation, notice that models relying on no-arbitrage in order to propagate liquidity shocks through a small number of factors (parallel shift, slope, convexity) usually predict that trading a low-liquidity asset is not expensive, as long as it is exposed to a liquid factor. In our framework, trading a low-liquidity asset is still expensive, which limits the ability to close arbitrage opportunities. This is because the empirical correlation matrices are never exactly low rank, as assumed in an idealized factor model, and heterogeneity in liquidity amplifies impact on directions of low volatility (for e.g. spread between tenor-matched cash bonds and futures).

**VII. CONCLUSION**

Prices at a given time are actually influenced by the history of all previous trades through a complex process that can be formalized within the *propagator model*. As



(a) Kyle matrix in relative price changes in basis point ( $10^{-4}$  of the asset price) by 100 millions USD worth of contract traded.



(b) Kyle matrix in absolute variations of annual yield in basis point ( $10^{-4}$  of the asset price) by 100 millions USD worth in equivalent 10-year bond.

FIG. 19: Kyle matrix  $\Lambda_{\text{Kyle}}$  on 9 November 2021 for a bin size of 30 minutes. Units are chosen to represent either relative price changes (Fig. 19a), or absolute variations of annual yield (Fig. 19b).

its calibration is computationally intensive, our study of multidimensional price formation focuses on linear models. While the auto-correlation of signed order flows invalidates these models, they remain significant to predict prices. Notably, we have demonstrated that accurate predictions of price variations can be achieved by appropriately considering the time scale, the correlation among assets, and the liquidity, while increasing the number of explanatory assets. More importantly, we have shown that highly liquid assets determine their prices internally and that their trades influence the prices of correlated, less liquid assets. In the case of interest rate markets, the 10-year bond future serves as the main liquidity reservoir influencing the prices of the other tenors, contrary to prevailing Financial Economics theories.

However, our analysis has revealed certain gaps. Certain asset prices are best explained by their trades at significantly longer time scales than suggested by their trading frequency. More generally, our investigation into the sensitivity of optimal cross-impact time scales to asset characteristics has identified two distinct groups of asset pairs in multiple cases. Further research could explore the factors that differentiate these groups.

## VIII. ACKNOWLEDGMENTS

We would like to express our gratitude to Mehdi Tomas, Jean-Philippe Bouchaud, and Natasha Hey, who contributed to our research through fruitful discussions. We are also indebted to Bertrand Hassani, who provided us with the opportunity to conduct this study at Quant AI Lab. We extend our appreciation to Emmanuel Serie and Stephen Hardiman for their assistance in maximizing the potential of the high-performance computing tools which were instrumental in completing this study. Finally, we would like to thank Cécilia Aubrun for her helpful proofreading.

This research was conducted within the Econophysics & Complex Systems Research Chair, under the aegis of the Fondation du Risque, the Fondation de l'École polytechnique, the École polytechnique and Capital Fund Management.

## REFERENCES

- A. S. Kyle, *Econometrica* **53**, 1315 (1985).  
 T. F. Loeb, *Financial Analysts Journal* **39**, 39 (1983).  
 V. Plerou, H. E. Stanley, X. Gabaix, and P. Gopikrishnan, *Quantitative Finance* **4**, 11 (2004).  
 R. Almgren, C. Thum, E. Hauptmann, and H. Li, *Risk* **18**, 58 (2005).  
 R. Kissell and R. Malamut, *The Journal of Trading* **1**, 12 (2005).  
 E. Moro, J. Vicente, L. G. Moyano, A. Gerig, J. D. Farmer, G. Vaglica, F. Lillo, and R. N. Mantegna, *Physical Review E* **80**, 066102 (2009).  
 B. Toth, Y. Lempriere, C. Deremble, J. de Lataillade, J. Kockelkoren, and J.-P. Bouchaud, *Physical Review X* **1**, 021006 (2011).  
 I. Mastromatteo, B. Toth, and J.-P. Bouchaud, *Physical Review E* **89**, 042805 (2014).  
 J. Donier and J. Bonart, *Market Microstructure and Liquidity* **01**, 1550008 (2015).  
 E. Zarinelli, M. Treccani, J. D. Farmer, and F. Lillo, *Market Microstructure and Liquidity* **01**, 1550004 (2015).  
 A. S. Kyle and A. A. Obizhaeva, *Review of Finance*, rfa008 (2023).  
 E. Bacry, A. Iuga, M. Lasnier, and C.-A. Lehalle, *Market Microstructure and Liquidity* **01**, 1550009 (2015).  
 B. Tóth, Z. Eisler, and J.-P. Bouchaud, *Market Microstructure and Liquidity* **03**, 1850002 (2017).  
 J.-P. Bouchaud, J. Bonart, J. Donier, and M. Gould, *Trades, Quotes and Prices: Financial Markets Under the Microscope* (Cambridge University Press, Cambridge, UK ; New York, 2018).  
 J. Hasbrouck, *The Journal of Finance* **46**, 179 (1991).  
 Z. Chen, W. Stanzl, and M. Watanabe, *Price Impact Costs and the Limit of Arbitrage*, Yale School of Management Working Papers ysm251 (Yale School of Management, 2002).  
 F. Lillo, J. D. Farmer, and R. N. Mantegna, *Nature* **421**, 129 (2003).  
 M. Potters and J.-P. Bouchaud, *Physica A: Statistical Mechanics and its Applications* **324**, 133 (2003).  
 W.-X. Zhou, *Quantitative Finance* **12**, 1253 (2012).  
 P. Gomber, U. Schweickert, and E. Theissen, *European Financial Management* **21**, 52 (2015).  
 D. E. Taranto, G. Bormetti, and F. Lillo, *Journal of Statistical Mechanics: Theory and Experiment* **2014**, P06002 (2014).  
 A. Kempf and O. Korn, *Journal of Financial Markets* **2**, 29 (1999).  
 V. Plerou, P. Gopikrishnan, X. Gabaix, and H. E. Stanley, *Physical Review E* **66**, 027104 (2002).  
 M. D. D. Evans and R. K. Lyons, *Journal of Political Economy* **110**, 170 (2002).  
 T. Chordia and A. Subrahmanyam, *Journal of Financial Economics* **72**, 485 (2004).  
 X. Gabaix, P. Gopikrishnan, V. Plerou, and H. E. Stanley, *The Quarterly Journal of Economics* **121**, 461 (2006).  
 C. Hopman, *Quantitative Finance* **7**, 37 (2007).  
 F. Patzelt and J.-P. Bouchaud, *Journal of Statistical Mechanics: Theory and Experiment* **2017**, 123404 (2017).  
 F. Patzelt and J.-P. Bouchaud, *Physical Review E* **97**, 012304 (2018).  
 J.-P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart, *Quantitative Finance* **4**, 176 (2004).  
 J.-P. Bouchaud, J. Kockelkoren, and M. Potters, *Quantitative Finance* **6**, 115 (2006).  
 J.-P. Bouchaud, J. D. Farmer, and F. Lillo, in *Handbook of Financial Markets: Dynamics and Evolution* (arXiv, 2009).  
 J. Gatheral, *Quantitative Finance* **10**, 749 (2010).  
 J. Gatheral and A. Schied, in *Handbook on Systemic Risk*, edited by J.-P. Fouque and J. A. Langsam (Cambridge University Press, 2013) 1st ed., pp. 579–602.  
 A. Alfonsi, F. Klöck, and A. Schied, *Mathematics of Operations Research* **41**, 914 (2016).  
 N. Gârleanu and L. H. Pedersen, *Journal of Economic Theory* **165**, 487 (2016).  
 D. E. Taranto, G. Bormetti, J.-P. Bouchaud, F. Lillo, and B. Tóth, *Quantitative Finance* **18**, 903 (2018).

- I. Ekren and J. Muhle-Karbe, *Mathematical Finance* **29**, 1066 (2019).
- J. Hasbrouck and D. J. Seppi, *Journal of Financial Economics* **59**, 383 (2001).
- T. Chordia, A. Subrahmanyam, and V. Anshuman, *Journal of Financial Economics* **59**, 3 (2001).
- M. D. D. Evans and R. K. Lyons, *Why Order Flow Explains Exchange Rates*, Tech. Rep. (University of California, 2001).
- J. Harford and A. Kaul, *Journal of Financial and Quantitative Analysis* **40**, 29 (2005).
- P. Pasquariello and C. Vega, *Review of Financial Studies* **20**, 1975 (2007).
- S. C. Andrade, C. Chang, and M. S. Seasholes, *Journal of Financial Economics* **88**, 406 (2008).
- H. E. Tookes, *The Journal of Finance* **63**, 379 (2008).
- P. Pasquariello and C. Vega, *Review of Finance* **19**, 229 (2015).
- S. Wang and T. Guhr, *Market Microstructure and Liquidity* **03**, 1850009 (2017).
- M. Benzaquen, I. Mastromatteo, Z. Eisler, and J.-P. Bouchaud, *Journal of Statistical Mechanics: Theory and Experiment* **2017**, 023406 (2017).
- M. Schneider and F. Lillo, *Quantitative Finance* **19**, 137 (2019).
- M. Tomas, I. Mastromatteo, and M. Benzaquen, Cross impact in derivative markets (2021), arXiv:2102.02834.
- M. Tomas, I. Mastromatteo, and M. Benzaquen, *Quantitative Finance* **22**, 1017 (2022).
- D. Brigo, F. Graceffa, and E. Neuman, *Quantitative Finance* **22**, 171 (2022).
- M. Schneider, *Market Microstructure, Price Impact and Liquidity in Fixed Income Markets*, Ph.D. thesis, Scuola Normale Superiore Pisa (2019).
- S. Wang, Trading strategies for stock pairs regarding to the cross-impact cost (2017).
- G. Zumbach and P. Lynch, *Physica A: Statistical Mechanics and its Applications* **298**, 521 (2001).
- P. E. Lynch and G. O. Zumbach, *Quantitative Finance* **3**, 320 (2003).
- M. Cordi, D. Challet, and S. Kassibrakis, *Quantitative Finance* **21**, 295 (2021).
- M. Rosenbaum and M. Tomas, *Frontiers of Mathematical Finance* **1**, 491 (2022).
- F. Cordon, C. Giannetti, F. Lillo, and G. Bottazzi, *Simulation*, 003754972211389 (2022).
- J. P. Bouchaud, Price Impact (2009), arXiv:0903.2428 [q-fin].
- T. W. Epps, *Journal of the American Statistical Association* **74**, 291 (1979).
- B. Toth and J. Kertesz, *Quantitative Finance* **9**, 793 (2009).
- W. S. Cleveland, *Journal of the American Statistical Association* **74**, 829 (1979).
- R. Renò, *International Journal of Theoretical and Applied Finance* **06**, 87 (2003).
- H. Theil, in *Henri Theil's Contributions to Economics and Econometrics*, Vol. 23, edited by A. J. H. Hallet, J. Marquez, B. Raj, and J. Koerts (Springer Netherlands, Dordrecht, 1992) pp. 345–381.
- P. K. Sen, *Journal of the American Statistical Association* **63**, 1379 (1968).
- A. F. Siegel, *Biometrika* **69**, 242 (1982).
- D. Brigo and F. Mercurio, *Interest Rate Models — Theory and Practice: With Smile, Inflation and Credit* (Springer International Publishing, 2006).
- W. Krämer, *Economics Letters* **30**, 37 (1989).
- G. E. P. Box and G. S. Watson, *Biometrika* **49**, 93 (1962).
- J. G. MacKinnon and H. White, *Journal of Econometrics* **29**, 305 (1985).
- J. S. Long and L. H. Ervin, *The American Statistician* **54**, 217 (2000).
- J. J. Goeman and A. Solari, *Statistics in Medicine* **33**, 1946 (2014).
- A. V. Frane, *Journal of Modern Applied Statistical Methods* **14**, 12 (2015).
- Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289 (1995).
- F. Lillo and J. D. Farmer, *Studies in Nonlinear Dynamics & Econometrics* **8**, 10.2202/1558-3708.1226 (2004).

## Appendix A: Notations

| Expression                       | Definition   |
|----------------------------------|--|
| $n$                              | The number of assets.  |
| $\mathcal{M}_n(\mathbb{R})$      | The set of real-valued square matrices of dimension $n$ .  |
| $\mathcal{O}_n$                  | The set of orthogonal matrices.  |
| $\mathcal{S}_n^+(\mathbb{R})$    | The set of real symmetric positive semi-definite matrices.   |
| $\mathcal{S}_n^{++}(\mathbb{R})$ | The set of real symmetric positive definite matrices.  |
| $A$                              | A matrix.  |
| $A^\top$                         | The transpose of matrix $A$ .  |
| $A^{1/2}$                        | A matrix such that $A^{1/2}(A^{1/2})^\top = A$ .   |
| $\sqrt{A}$                       | The unique positive semi-definite symmetric matrix such that $(\sqrt{A})^2 = A$ .                                |
| $\text{diag}(A)$                 | The vector in $\mathbb{R}^n$ formed by the diagonal items of $A$ .   |
| $\text{diag}(v)$                 | The diagonal matrix whose components are the components $(v_1, \dots, v_n)$ of $v \in \mathbb{R}^n$ .            |
| $\tau$                           | The bin size.  |
| $p_{t,i}$                        | The opening price of asset $i$ in the time window $[t, t + \tau]$ .  |
| $p_t$                            | The vector of asset prices at opening in the time window $[t, t + \tau]$ .                                       |
| $q_{t,i}$                        | The net market order flow traded during the time window $[t, t + \tau]$ .  |
| $q_t$                            | The vector of the net traded order flows during the time window $[t, t + \tau]$ .                                |
| $\Delta p_t$                     | the prices changes $p_{t+\tau} - p_t$ during the time window $[t, t + \tau]$ .                                   |
| $\Lambda_t$                      | The cross-impact matrix at time $t$ .  |
| $\eta_t$                         | The vector of zero-mean random variables representing exogenous noise at time $t$ .                              |
| $\Sigma_t$                       | The price change covariance matrix at time $t$ .   |
| $\Omega_t$                       | The order flow covariance matrix at time $t$ .   |
| $R_t$                            | The response matrix between price variations and order flows at time $t$ .                                       |
| $\sigma_t$                       | The vector of price variation volatility at time $t$ .   |
| $\omega_t$                       | The vector of the signed order flow volatility at time $t$ .   |
| $\mathcal{R}^2(M)$               | The $M$ -weighted generalized R-squared.   |
| $\Delta \mathcal{R}^2(M)$        | The accuracy increase from the cross sectional model.  |
| $\mathcal{R}^{2*}(M)$            | The maximum goodness-of-fit observed empirically across the tested bin size $\tau$ .                             |
| $\tau^*(M)$                      | The optimal time scale corresponding to the maximum goodness-of-fit $\mathcal{R}^{2*}(M)$ .                      |
| $\Delta \mathcal{R}^{2*}(M)$     | The maximum accuracy increase $\Delta \mathcal{R}^2(M)$ observed empirically across the tested bin size $\tau$ . |
| $\tau_\Delta^*(M)$               | The optimal time scale corresponding to the maximum accuracy increase $\Delta \mathcal{R}^{2*}(M)$ .             |
| $f_i$                            | The trading frequency of the predicted asset $i$ .   |
| $f_j$                            | The trading frequency of the explanatory asset $j$ .   |
| $\rho_{ij}$                      | Price increments correlation between the assets $i$ and $j$ .  |
| $\bar{\sigma}_i$                 | The average across time of the price variation volatility of asset $i$ .   |
| $\bar{\omega}_i$                 | The average across time of the signed order flow volatility of asset $i$ .                                       |
| $\bar{\omega}_i \bar{\sigma}_i$  | The liquidity of the predicted asset $i$ .   |
| $\bar{\omega}_j \bar{\sigma}_j$  | The liquidity of the explanatory asset $j$ .   |



## Appendix B: Statistical significance analysis

### 1. Statistical significance of the Goodness-of-fit

The main results of our analysis are expressed in terms of a generalized  $\mathcal{R}^2(M)$ . For a single asset  $i$ , the indicator  $\mathcal{R}^2(I_{\sigma_i})$  is precisely the R-squared of the linear regression of its price increments over its predicted price increments in the model with no Y-ratio:

$$\Delta p_{t,i} = Y \widehat{\Delta p}_{t,i} + \eta_{t,i}, \quad (\text{B1})$$

where the explanatory variable  $\widehat{\Delta p}_{t,i}$  is the prediction of the model with no Y-ratio.

The significance of the R-squared of the above regression can be provided by an F-test. Indeed, this latter allows us to compare two models, one model being the reduction of the other to fewer parameters. Here we compare the model with one explanatory variable to the model with only an intercept. Let  $\tilde{\xi} \in \mathbb{R}^N$  denote the vector of the errors estimated in the model with no explanatory variable, and  $\hat{\xi} \in \mathbb{R}^N$  denote the vector of the errors estimated in the cross-impact model. The F-statistic is expressed as the normalized difference between the squared errors in the two models:

$$F = \frac{\tilde{\xi}^\top \tilde{\xi} - \hat{\xi}^\top \hat{\xi}}{\hat{\xi}^\top \hat{\xi}}. \quad (\text{B2})$$

Since the errors in the parameter-free model are precisely equal to the centered explained variable, denoted as  $y \in \mathbb{R}^N$ , we can express this F-statistics as a function of the  $\mathcal{R}^2$ :

$$F = \frac{y^\top y - \hat{\xi}^\top \hat{\xi}}{\hat{\xi}^\top \hat{\xi}} = \frac{\hat{y}^\top \hat{y}}{\hat{\xi}^\top \hat{\xi}} = \frac{\mathcal{R}^2}{1 - \mathcal{R}^2}. \quad (\text{B3})$$

Under the usual assumptions for the residuals and the explained variable, this F-statistics follows a Fisher law  $F(N, N - 1)$ . The significance of the  $\mathcal{R}^2$  can be then provided by the p-values of the F-statistics of the linear model calibrating the Y-ratio.

Yet, the residuals in the above regression are auto-correlated due to the properties of the order flows. They are also non-Gaussian (heavy tails, negative skew) and heteroskedastic, due to the properties for the price process. In fact, returns are generally conditionally heteroskedastic but unconditionally homoskedastic. Here the unconditional heteroskedasticity observed in the data might be due to some trend in the annual sample.

However, the observed levels of auto-correlation are sufficiently low (around 10% at the lag 1) to avoid compromising the robustness of the test (Krämer, 1989). This issue is further studied in the following section. Moreover, the non-Gaussianity only partially limits the robustness of the F-statistic test (Box and Watson, 1962). Yet, the heteroskedasticity issue requires using a modified F-statistic test robust to this assumption. Thus, we

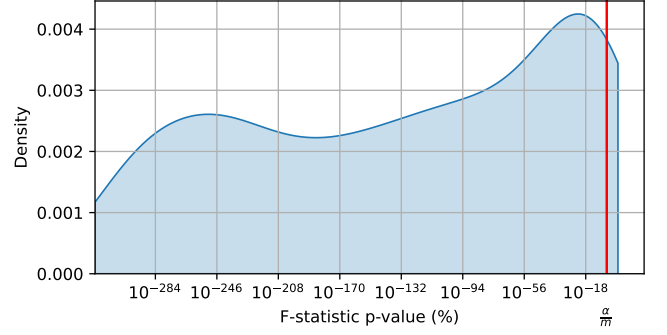


FIG. 20: Empirical distribution of the statistical significance of the  $\mathcal{R}^{2*}(I_{\sigma_i})$  across the years and assets in our sample.

measure the statistical significance of the  $\mathcal{R}^2$  using the approach of MacKinnon and White (1985) (implemented in the *Statsmodels* python library through the method of Long and Ervin (2000)).

These F-statistics confirm that the  $\mathcal{R}^{2*}$  displayed in our study are significant. Specifically, in the single asset case, each optimal goodness-of-fit  $\mathcal{R}^{2*}$  is obtained from a linear regression. The F-statistics p-values of  $m \approx 10^3$  (500 assets across 5 years) statistical tests are exhibited in Fig. 20. Notably, only few p-values are above the Bonferroni upper bound (Goeman and Solari, 2014; Frane, 2015). This upper bound is used for the identification of false positive when performing multiple hypothesis tests. Here, the rate of false positive at the confidence interval  $\alpha = 10^{-2}$  is bounded by the share of p-values above  $\frac{\alpha}{m} = 10^{-5}$ . Fig. 20 shows that only a negligible share of these p-values are not significant (approximately 2%). The more accurate procedure from Benjamini and Hochberg (1995) yields similar results. If we compare the p-value of rank  $k$  (in ascending order) to  $\frac{k\alpha}{m}$ , we find that 1.8% of these p-values are above this threshold.

### 2. Auto-correlation structure and comparison with the propagator model

The auto-correlation of signed order flows is a well-documented feature of financial markets (Lillo and Farmer, 2004). Using the E-mini S&P future binned every 1 minute to calibrate the single asset model, we observe significant auto-correlation of both the signed order flows and the residuals (Fig. 21). In contrast, the auto-correlation of prices is of the same size as the noise. If the number of data points increases, prices will become even more efficient, so the auto-correlation of the residuals will increase to compensate for the long memory of the signed order flows. Thus, the model will be invalidated.

As previously mentioned, one approach to re-conciliate the long memory of the order flows with the efficiency of

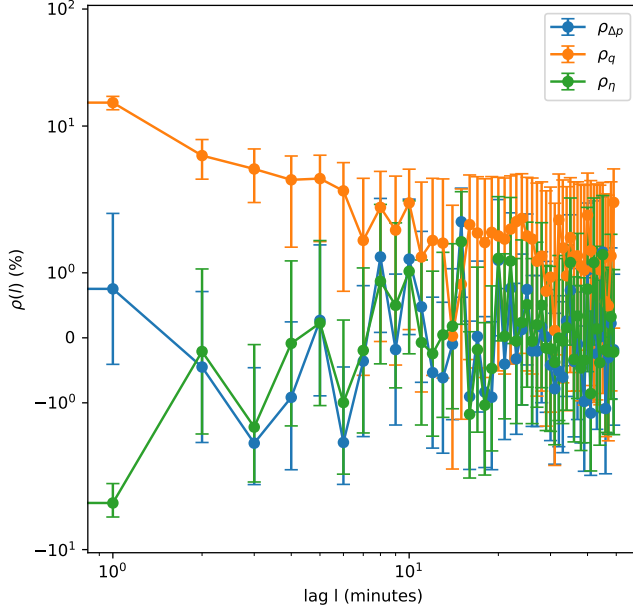


FIG. 21: Auto-correlation of price variations  $\Delta p$ , signed order flows  $q$  and residuals  $\eta$  for the E-mini S&P future. Data is binned every 1 minute for the year 2021. Error bars represent one standard deviation confidence interval.

prices is to define a *propagator model* (Bouchaud *et al.*, 2018, 2006; Alfonsi *et al.*, 2016; Benzaquen *et al.*, 2017; Schneider and Lillo, 2019; Bouchaud, 2009) as follow:

$$p_t = \sum_{s \leq t} G(t-s)q_s + \eta_t, \quad (\text{B4})$$

where  $G : t \rightarrow G(t) \in \mathcal{M}_n(\mathbb{R})$  captures the dependence on past order flows and  $\eta_t$  is a vector of zero-mean random variables. As shown by Tomas *et al.* (2022) the calibration of the true propagator model would yield only marginal improvements in the goodness-of-fit. However, this model is significantly more complex to calibrate, which would impede conducting this study at the same scale across time and assets.