The European Journal of Finance

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rejf20

A Hawkes process analysis of high-frequency price endogeneity and market efficiency

Jingbin Zhuo, Yufan Chen, Bang Zhou, Baiming Lang, Lan Wu & Ruixun Zhang

To cite this article: Jingbin Zhuo, Yufan Chen, Bang Zhou, Baiming Lang, Lan Wu & Ruixun Zhang (04 Sep 2023): A Hawkes process analysis of high-frequency price endogeneity and market efficiency, The European Journal of Finance, DOI: 10.1080/1351847X.2023.2251531

To link to this article: https://doi.org/10.1080/1351847X.2023.2251531

THE EUROPEAN JOURNAL OF FINANCE
https://doi.org/10.1080/1351847X.2023.2251531

Routledge
Taylor & Francis Group

A Hawkes process analysis of high-frequency price endogeneity and market efficiency

Jingbin Zhuo [a], Yufan Chen [a], Bang Zhou [a], Baiming Lang [a], Lan Wu [a,b,c] and Ruixun Zhang [a,b,c,d]

[a]Department of Financial Mathematics, School of Mathematical Sciences, Peking University, Beijing, People's Republic of China; [b]Laboratory for Mathematical Economics and Quantitative Finance, Peking University, Beijing, People's Republic of China; [c]Center for Statistical Science, Peking University, Beijing, People's Republic of China; [d]National Engineering Laboratory for Big Data Analysis and Applications, Beijing, People's Republic of China

ABSTRACT

We use the Hawkes process to model the high-frequency price process of 108 stocks in the Chinese stock market, in order to understand the endogeneity of price changes and the mechanism of information processing. Using a piece-wise constant exogenous intensity, we employ non-parametric estimation, residual analysis, and Bayesian Information Criterion (BIC) to determine that a power-law kernel is the most appropriate for our data. We propose the internal branching ratio to represent endogeneity within a finite interval. The branching ratio tends to be higher after the market opens and before the market closes, with a mean value of around 0.81, suggesting significant endogeneity in price changes. In addition, we explore the relationship between branching ratios and stock characteristics using panel regression. Higher branching ratios are associated with lower levels of price efficiency at high, but not low, frequencies. Finally, the branching ratio increases over time without significant impact from COVID-19.

ARTICLE HISTORY
Received 1 November 2022
Accepted 2 August 2023

KEYWORDS
Market efficiency; price endogeneity; Hawkes process; branching ratio; Kernel function; Chinese stock market

1. Introduction

The debate over the efficiency of financial markets has generated a great deal of attention, particularly regarding the 'endo-exo problem' (Wheatley, Wehrli, and Sornette 2019), which concerns the decomposition of market activities into endogenous and exogenous parts. The efficient market hypothesis (EMH) posits that the market fully and instantaneously reflects all available information in asset prices (Fama 1970; Samuelson 1965), such that market prices are driven solely by exogenous inputs of information. However, empirical studies have shown that a significant fraction of asset price fluctuations cannot be explained by changes in their underlying fundamental value (Cutler, Poterba, and Summers 1989; Fair 2002). To better understand the dynamics of financial markets, Lo (2004, 2017) proposes the Adaptive Markets Hypothesis to complement EMH, which maintains that the actual price process incorporates information gradually and adaptively.

In this article, we use the Hawkes process (Hawkes 1971a,b) to model the high-frequency price process in order to understand the endogeneity of price changes and the mechanism of information processing. Hawkes processes have emerged as a popular tool for studying fluctuations in high-frequency financial prices (Bacry, Mastromatteo, and Muzy 2015; Hawkes 2018, 2020), partly due to their ability to naturally decompose endogenous and exogenous components within a branching representation (Ogata 1981). By leveraging this powerful framework, researchers can better understand the process by which information is incorporated into prices, and the intricate relationships between different variables that influence information processing in financial markets.

CONTACT Ruixun Zhang zhangruixun@pku.edu.cn

A crucial debate regarding Hawkes processes centers around whether the observed processes are *critical*, which is equivalent to whether the branching ratio η is greater or less than one (see Section 2 for mathematical details of the Hawkes processes).[1] In particular, it has been shown that inadequate treatments of major features of the data, including trends (e.g. intraday seasonality), external shocks, and data artifacts (e.g. limited data resolution), can upward bias the estimates of the degree of self-excitation and memory (Filimonov and Sornette 2015; Wehrli, Wheatley, and Sornette 2021). Filimonov and Sornette (2015) apply Hawkes processes to the high-frequency data of E-mini S&P 500 futures contracts and study how to deal with trends, choose an appropriate kernel function, and analyze possible sources of estimation bias.

Our study uses high-frequency data from 108 liquid stocks on China's Shenzhen Stock Exchange that are constituents of the CSI 300 index from 2019 to 2020. Among the international markets, the Chinese financial market is of particular importance because of its growing size[2] as well as increasing centrality in the global financial system.[3] In addition, the Chinese stock market has several unique features that make research on market microstructure and the mechanism of information processing both interesting and challenging. First, unlike developed markets that are dominated by institutional investors, the Chinese stock market is dominated by retail investors (Jones et al. 2020; W. Li and Wang 2010). The speculative and short-term trading motives of many retail investors may lead to very different price formation mechanisms compared to institutional investor-driven markets. Second, short sales of individual stocks are very difficult to carry out in China.[4] Although there is no broad consensus, many academics agree that a lack of short-selling hinders price discovery, rendering markets less efficient (Saffi and Sigurdsson 2011). Third, the degree of automation in the Chinese market, though has increased in recent years, is still relatively low compared to developed markets such as the US.

We conduct a comprehensive analysis to show that the power-law kernel is the most appropriate parametric form of the kernel function for data in the Chinese stock market. We measure the level of endogeneity for stocks in the Shenzhen Stock Exchange with different lengths of time windows. We also analyze the level of endogeneity cross-sectionally and study its relationship to classical proxies of price efficiency as well as other stock characteristics. In addition, we check whether the endogeneity of price processes changes over time, and whether it is affected by COVID-19. To the best of our knowledge, this is the first study on the 'endo-exo problem' in the Chinese financial market.

First, we use non-parametric methods based on the Fourier transform to estimate the kernel function of a shock event in the market. We also compare the log-likelihood and residuals of various parametric kernel functions. Both analyzes suggest that the power-law kernel function of the form (4) is the most appropriate. We then set the exogenous intensity function as piece-wise constant and follow Wheatley, Wehrli, and Sornette (2019) to determine the optimal number of segments for the exogenous intensity function given different lengths of estimation window based on the Bayesian Information Criterion (BIC).

Next, to accommodate the fact that stock markets open for a limited time each day, we measure endogeneity by proposing the *internal branching ratio*. We introduce a maximum likelihood estimation method to aggregate data across different trading days and estimate the internal branching ratio for different time windows within a day. We observe that the branching ratio tends to be higher after the market opens and before the market closes. The mean value of the branching ratio is around 0.79, and its aggregate estimate is around 0.81. This implies significant endogeneity in price changes, although not reaching criticality.

Finally, we estimate and compare the branching ratio for 108 stocks cross-sectionally, and study the impact of the COVID-19 pandemic on the branching ratio. The branching ratio measures the ability of the market to absorb information, thus intuitively it should be related to market efficiency. Our empirical results show that higher branching ratios are associated with lower levels of price efficiency, as measured by the variance ratio statistic (Lo and MacKinlay 1988), at high frequencies. However, this phenomenon disappears at low frequencies, suggesting that activities from high-frequency traders are driving this phenomenon. Moreover, the branching ratio of the Chinese stock market is increasing over time, and is not significantly influenced by the COVID-19 pandemic.

The remainder of the article is organized as follows. In Section 2, we introduce the Hawkes process and propose a new measure, the internal branching ratio, to measure the endogeneity of a sample within a finite window. Section 3 describes the data. Section 4 discusses model selection. Section 5 reports estimation results of the internal branching ratio in the Chinese stock market and provides its economic interpretation.

Section 6 provides several robustness checks. Section 7 concludes. We provide additional technical details in the Appendix.

2. Hawkes processes

The methodology for estimating the endogeneity in the dynamics of a given point process is based on the self-excited conditional Poisson model introduced by Hawkes (1971a,b). As a point process, given an ordered set of event times, $\{t_i : i = 1, 2, \ldots\}$, satisfying $t_i \leqslant t_j$ for $i < j$, the Hawkes process is the corresponding counting process, $N(t) = \max\{i : t_i \geqslant t\}$. Moreover, a point process is fully characterized by its conditional intensity:

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{h \downarrow 0} \frac{1}{h} \mathbb{P}\left[N(t+h) - N(t) > 0 | \mathcal{F}_{t-}\right], \tag{1}$$

where $\mathcal{F}_{t-} = \{t_1, \ldots, t_i : t_i < t\}$ is the filtration that represents the history of the process until time t. For Hawkes processes, the conditional intensity takes the following general form:

$$\lambda(t) \equiv \mu(t) + \int_{-\infty}^{t} \phi(t-s) \mathrm{d}N_s. \tag{2}$$

Here $\mu(t)$ is referred to as the *exogenous intensity* (or background intensity), which is a deterministic function of time that accounts for the intensity of arrival of exogenous events that are independent of history, and the deterministic kernel function $\phi(t)$ models the *endogenous* feedback mechanism and captures the memory effects of the process.

Next, we present the parametric forms of the kernel function in Section 2.1, establish the definition of the branching ratio in Section 2.2, and elaborate on the estimation of the Hawkes process in Section 2.3. In Section 2.4, we propose the concept of *internal branching ratio* as a measure of endogeneity of a sample of finite length.

2.1. Parametric forms of the Kernel function

Hawkes processes typically use parametric kernel functions $\phi(t)$. In particular, we consider the following three candidates.

The **exponential kernel**, first proposed by Hawkes (1971b), is defined as

$$\phi(t|\eta, \beta) = \eta\beta e^{-\beta t} \tag{3}$$

with parameters η and β. This exponential form ensures the Markovian property (Oakes 1975) and is widely adopted in financial applications (Bowsher 2007; Cont 2011; Filimonov et al. 2014; Filimonov and Sornette 2012).

The **power-law kernel** is defined as

$$\phi(t|\eta, p, c) = \eta p c^p (t+c)^{-1-p}, \tag{4}$$

with parameters η, p and c. This power-law form comes from the application of Hawkes processes in geophysics (Helmstetter and Sornette 2002; Ogata 1988; Vere-Jones 1970; Vere-Jones and Ozaki 1982).

The **exponential power-law kernel**, recently adopted by Hardiman, Bercot, and Bouchaud (2013), represents the sum of exponential functions with a power-law decay. Specifically, it is given by

$$\phi(t|\eta, \varepsilon, \tau_0) = \frac{\eta}{Z} \left(\sum_{i=0}^{M-1} \left(\frac{1}{\xi_i} \right)^{1+\varepsilon} e^{-\frac{t}{\xi_i}} - S e^{-\frac{t}{\xi_{-1}}} \right) \tag{5}$$

where $\xi_i = \tau_0 m^i$, $-1 \leqslant i < M$, and Z and S are normalizing parameters that ensure $\int_0^\infty \phi(t)\mathrm{d}t = \eta$ and $\phi(0) = 0$. This kernel function allows for a delay in the shock of an event, resulting in a peak at a specific lag controlled by τ_0. Moreover, it approximately follows a power law at the tail.

2.2. Branching ratio

The branching ratio of Hawkes processes, including those specified by all three kernel functions in Section 2.1, is defined as

$$\eta \equiv \int_0^\infty \phi(t)\mathrm{d}t. \tag{6}$$

We provide a few remarks to illustrate its intuition.

The definition of the branching ratio (6) is derived from the generalized branching process representation of Hawkes processes, which was initially introduced by Hawkes and Oakes (1974). According to this representation, a Hawkes process can be viewed as an immigrant-birth process, consisting of a Poisson immigrant with a rate of $\mu(t)$ and Poisson descendants with a rate of $\phi(t)$.

In this context, the branching ratio η precisely represents the expected number of descendants triggered by an immigrant. Depending on the branching ratio, there are three regimes: (i) sub-critical ($\eta < 1$), (ii) critical ($\eta = 1$), and (iii) super-critical or explosive ($\eta > 1$). In the sub-critical and critical regimes, the process eventually dies out with a probability of 1, while in the super-critical regime, there exists a finite probability for the process to escalate to an infinite number of events.

Furthermore, in the case of a constant exogenous intensity ($\mu(t)$ is a constant) and in the sub-critical regime ($\eta < 1$), the branching ratio is exactly equal to the average fraction of the number of descendants in the whole population of events (Filimonov and Sornette 2012; Helmstetter and Sornette 2002). In other words, the branching ratio represents the average proportion of endogenously generated events relative to all events. Hence, the branching ratio serves as an effective measure of the system's level of endogeneity.

2.3. Parameter estimation

When applying Hawkes processes in practice, events are usually observed at times within a finite time interval, $[0, T]$:

$$0 \le t_1 < t_2 < \cdots < t_{N_T} \le T, \tag{7}$$

where N_T is the total number of the events occured in $[0, T]$. Therefore, we use maximum likelihood estimation and estimate parameters of the Hawkes process by maximizing the following log-likelihood function:

$$\begin{aligned} \ln L(t_1, t_2, \ldots, t_n|\boldsymbol{\theta}) &= -\int_0^T \lambda(t|\boldsymbol{\theta})\mathrm{d}t + \int_0^T \ln\lambda(t|\boldsymbol{\theta})\mathrm{d}N_t \\ &= -\int_0^T \lambda(t|\boldsymbol{\theta})\mathrm{d}t + \sum_{i=1}^{N_T} \ln\lambda(t_i|\boldsymbol{\theta}). \end{aligned} \tag{8}$$

In particular, we maximize the log-likelihood function over the set of parameters $\boldsymbol{\theta}$ using the L-BFGS-B algorithm (Byrd et al. 1995; Zhu et al. 1997). Generally, the calculation of the log-likelihood function (8) has a computational complexity of $\mathcal{O}(N_T^2)$. However, for exponential kernels (3) and exponential power-law (5) kernels, the complexity can be reduced to $\mathcal{O}(N_T)$ by leveraging a recursive relation (Ozaki 1979). For the power-law kernel (4), a discretization algorithm can be employed to achieve the same computational complexity reduction (Ogata, Matsuúra, and Katsura 1993). Additional details of the discretization algorithm for the power-law kernel can be found in Appendix 1.

Notably, optimizing the log-likelihood function (8) involves several challenges. The log-likelihood function is not convex when the kernel is specified by a power law or an exponential power law function. However, as demonstrated later in Section 6.3, despite its non-convex nature, the log-likelihood usually exhibits a unique global maximum under the power law in practice. This global maximum can be efficiently obtained through the L-BFGS-B algorithm. In contrast, when the kernel follows an exponential power law, the log-likelihood has two local maximums. Therefore, it is important to choose appropriate initial values for the optimization algorithm to converge to the global maximum. The specific details of our optimization are provided in Section 6.3.

After the parameters of the Hawkes process are estimated, we perform model evaluation by comparing with non-parametric estimation, using Bayesian information criterion (BIC), and conducting residual analysis. First, we employ non-parametric estimation of the Hawkes model as a benchmark for different forms of the background intensity and kernel function. The details of the non-parametric estimation are discussed in Section 4.1 and elaborated further in Appendix 2. Second, a crucial criterion for model selection is the Bayesian information criterion (BIC), defined as

$$\mathrm{BIC} = n_p \ln N_T - 2 \ln L, \tag{9}$$

where n_p denotes the number of parameters in the model, N_T represents the total number of events within the time window $[0, T]$, and $\ln L$ corresponds to the log-likelihood function (8). Finally, we validate our model with the goodness-of-fit using residual analysis (Ogata 1988). Specifically, we consider $\xi_i = \int_0^{t_i} \lambda(t)\mathrm{d}t$. If the estimation of $\lambda(t)$ is accurate, $\{\xi_i\}$ must be a Poisson process with a constant intensity of 1. To evaluate the performance of the three kernel functions, we calculate the sequence $\{\xi_i\}$ under each kernel. By comparing the deviations in the $\{\xi_i\}$ sequence with the exponential distribution, we can draw conclusions regarding the effectiveness of each kernel function.

2.4. Internal branching ratio

When applying Hawkes processes, the observed jump events are typically limited to a finite time interval $[0, T]$. This is true for most financial markets and, in particular, the Chinese stock market trades for four hours on each trading day. Therefore, the data of new events derived from observed events within the interval $[0, T]$ is truncated. When the kernel function decays slowly, such as in the case of a power-law kernel with a small exponent, the expected number of derived events for an event in the interval is much smaller than the branching ratio. This observation can be further supported by the following Proposition 2.1 which provides an expression of the estimated branching ratio.

First, we represent the window length, the kernel function, the branching ratio, and the normalization of the kernel function by T, $\phi(t)$, η, and $h(t)$ respectively, where $h(t) = \phi(t)/\eta$. To understand the relationship between events, we consider $\pi_{i,j} = \phi(t_i - t_j)/\lambda(t_i)$, which represents the probability that the ith event originates from the jth event. Furthermore, we define $H(t) = \int_0^t h(s)\mathrm{d}s$ and $H_0 = \sum_{i=1}^{N_T} H(T - t_i)$. Finally, we denote the estimated parameters by $\hat{\boldsymbol{\theta}}$ and calculate $\hat{\eta}$, $\hat{\pi}_{i,j}$, and $\hat{H}_0$ by substituting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is the parameter vector of the Hawkes model. We have the following proposition, which expresses the relationship between the above estimators.

Proposition 2.1: *The estimated parameters, $\hat{\eta}$, $\hat{\pi}_{i,j}$, and $\hat{H}_0$, satisfy the following equation:*

$$\hat{\eta} = \frac{\sum_{1 \leqslant j < i \leqslant N_T} \hat{\pi}_{i,j}}{\hat{H}_0}. \tag{10}$$

Proof: The log-likelihood satisfies:

$$\begin{aligned}
\ln L &= -\int_0^T \left(\mu(t) + \eta \sum_{t_i < t} h(t - t_i)\right) \mathrm{d}t + \sum_{i=1}^{N_T} \ln \left(\mu(t_i) + \eta \sum_{t_j < t_i} h(t_i - t_j)\right) \\
&= -\int_0^T \mu(t)\mathrm{d}t - \eta \sum_{i=1}^{N_T} \int_{t_i}^T h(t - t_i)\mathrm{d}t + \sum_{i=1}^{N_T} \ln \left(\mu(t_i) + \eta \sum_{t_j < t_i} h(t_i - t_j)\right).
\end{aligned}$$

Hence we have

$$\frac{\partial \ln L}{\partial \eta} = -H_0 + \sum_{i=1}^{N_T} \frac{\sum_{t_j < t_i} h(t_i - t_j)}{\mu(t_i) + \eta \sum_{t_j < t_i} h(t_i - t_j)} = -H_0 + \frac{\sum_{1 \leqslant j < i \leqslant N_T} \pi_{i,j}}{\eta}.$$

When $\ln L$ reaches its maximum at $\hat{\boldsymbol{\theta}}$, we have $\frac{\partial \ln L}{\partial \eta} = 0$, resulting in

$$\hat{\eta} = \frac{\sum_{1 \leqslant j < i \leqslant N_T} \hat{\pi}_{i,j}}{\hat{H}_0}.$$

■

Notably, $H(T - t_i)$ represents the expected proportion of offspring within the window $[0, T]$ to all offspring of the ith event and H_0 is the sum of these proportions. Therefore, the maximum likelihood estimation $\hat{\eta}$ extends the information within the finite window to infinite time. As a result, this measure of endogeneity can substantially differ from the actual proportion of endogenous events in the interval. Furthermore, as shown empirically in Section 4.1, the actual kernel function decays with a power law. Prior research by Hardiman, Bercot, and Bouchaud (2013) highlights that a power-law decaying kernel may exhibit a lower exponent at short times and a higher exponent at long times. In this case, if the observed data only captures the part with a lower exponent, extending the kernel function potentially leads to a substantial error in estimating the branching ratio.

For these reasons, it is not appropriate to directly use the original definition of the branching ratio (6) to measure the endogeneity within a finite window. To address this problem, we propose the *internal branching ratio* to measure the endogeneity of a process within a window of finite length.

Definition 2.1 (Internal Branching Ratio): For a Hawkes process with a conditional intensity function (2), the internal branching ratio over a finite time interval $[0, T]$ is defined as

$$\eta^{\text{in}} = \frac{\sum_{1 \leqslant j < i \leqslant N_T} \pi_{i,j}}{N_T} \tag{11}$$

where N_T is the total number of the events in the window $[0, T]$.

The following result shows that the internal branching ratio η^{in} is never greater than one.

Proposition 2.2: *The internal branching ratio satisfies the following equation:*

$$\eta^{in} = 1 - \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\mu(t_i)}{\lambda(t_i)}. \tag{12}$$

***Proof*:** According to Definition 2.1 and Equation (2), we have

$$\begin{aligned}
\eta^{\text{in}} &= \frac{\sum_{1 \leqslant j < i \leqslant N_T} \pi_{i,j}}{N_T} \\
&= \frac{1}{N_T} \left(\sum_{i=1}^{N_T} \frac{\sum_{j=1}^{i-1} \phi(t_i - t_j)}{\lambda(t_i)} \right) \\
&= \frac{1}{N_T} \left(\sum_{i=1}^{N_T} \frac{\lambda(t_i) - \mu(t_i)}{\lambda(t_i)} \right) \\
&= 1 - \frac{1}{N_T} \sum_{i=1}^{N_T} \frac{\mu(t_i)}{\lambda(t_i)}.
\end{aligned}$$

■

This new definition performs well in finite sample, as shown in Section 5.1.

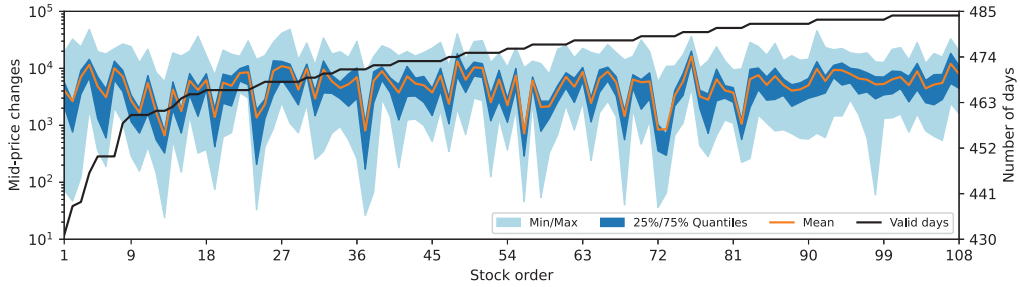


Figure 1. Descriptive statistics for all 108 stocks in the dataset. Stocks are sorted on the horizontal axis by the number of valid trading days. The black line shows the number of trading days that pass the screening criteria. The light blue shade shows the minimum and maximum number of daily mid-price changes across trading days for each stock. The blue shade shows the 25th and 75th percentile of the same quantity. The red line shows the mean value of the same quantity.

Table 1. Descriptive statistics for five sample stocks in the dataset. For each stock, we should the stock ID, the number of trading days that pass the screening criteria, and the minimum, 25th percentile, median, mean, 75th percentile, and maximum number of daily mid-price changes.

Stock ID	Valid dates	Minimum	25% quantile	Median	Mean	75% quantile	Maximum
000001	482	716	2157	3373	4060	5158	16,680
000002	483	610	3694	5338	5995	7312	23,131
000063	468	2502	6352	9092	10,930	13,243	41,308
000069	481	224	564	847	1063	1224	9782
000100	467	33	206	718	1369	1886	14,289

3. Data

3.1. Raw data

We obtain the raw data from the Shenzhen Stock Exchange (SZSE) Historical Tick Data of the Chinese stock market.[5] We select stocks that are constituents of the CSI 300 Index, which contains 121 liquid stocks in our sample period.[6] Our data contains tick-level trade and quote records over 484 trading days, starting from January 2nd, 2019 to December 31st, 2020. The timestamps in the raw data have a precision of 10 milliseconds. We exclude several trading days due to data corruption.[7] We also exclude stocks that are listed after December 1st, 2018, and stocks that are suspended for more than one month from 2019 to 2020. This leads to a final sample of 108 stocks for our analysis.

3.2. Preprocessing

Our research focuses on the mid-price process of individual stocks. We follow the data preprocessing in Zhang et al. (2022) to reconstruct the order book during continuous bidding.[8] We use the most granular observations of the mid-price process, whose timestamps have a precision of 10 milliseconds. Therefore, there can still exist multiple mid-price changes marked with the same timestamp, in which case we redistribute those events evenly across the 10-millisecond window.[9]

In order to ensure reliable parameter estimation, for each stock, we only include trading days in which the stock price does not reach the daily price limits, does not experience temporary suspensions from major events, and the mid-price jumps more than 20 times.

Table 1 provides the number of valid trading days and the descriptive statistics of the daily mid-price changes for five sample stocks, and Figure 1 provides the same statistics for all 108 stocks in our sample. The vast majority of stocks have an average of at least 1,000 mid-price changes per day. This allows us to model the price changes with the Hawkes process and estimate parameters reliably. Meanwhile, the vast majority of stocks have at least 95% of all trading days available in our analysis, which shows that the filtering should not lead to large selection biases for our analysis. The trading days mentioned hereafter are the valid trading days.

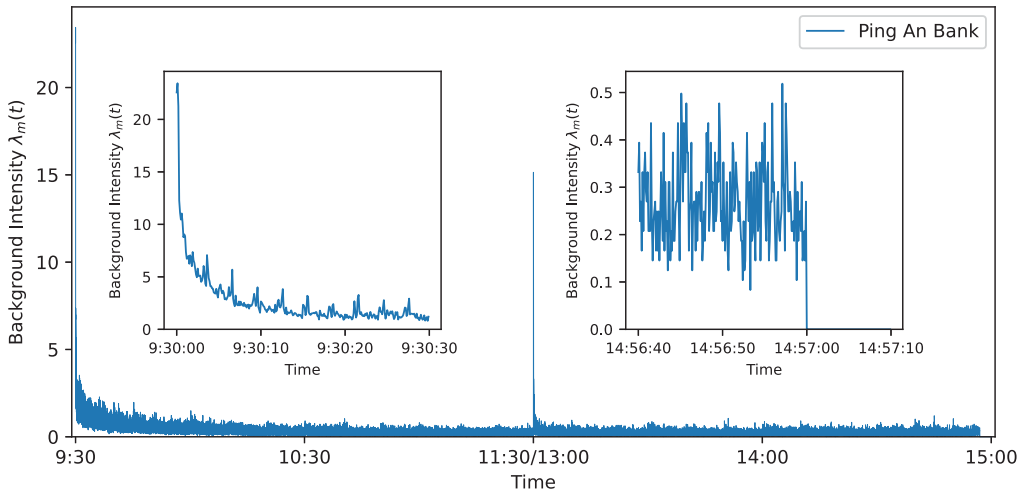


Figure 2. The main panel shows the background intensity $\lambda_m(t)$ of Ping An Bank. The empirical intensity function of each trading day is calculated with a step size of 0.1 seconds. The left sub-panel shows the intensity function for the first 30 seconds in the morning, while the right sub-panel shows the intensity function for the last 20 seconds before the closing auction and the first 10 seconds during the closing auction.

3.3. Detrending

The general sub-critical Hawkes process is stationary. However, it is well-known that trading activities exhibit strong intraday patterns and are considered non-stationary. For example, the mid-price changes rapidly at the start of each morning and afternoon session, which aligns with the observed liquidity patterns of the Chinese market (Zhang et al. 2022). Moreover, Wu, Zhang, and Dai (2022) and Wehrli, Wheatley, and Sornette (2021) report the existence of periodic patterns in mid-price changes in both the Chinese and US stock markets as well as EUR/USD data, respectively. As an illustrative example, Figure 2 demonstrates the background intensity function of the mid-price process for a sample stock, Ping An Bank, averaged over trading days.

To prepare the data for analysis with a Hawkes process, we follow the approach outlined by Hardiman, Bercot, and Bouchaud (2013) and Filimonov and Sornette (2015) to remove any underlying trends. Specifically, we assume that the timestamps t_i arise from a Hawkes process N_t with a known exogenous intensity function $\mu(t)$. Under this assumption, we calculate the corresponding timestamps r_i as $\int_0^{t_i} \mu(t)\mathrm{d}t$, which effectively removes any dependence on the original exogenous intensity function. As a result, we can treat the resulting timestamps approximately as observations from a Hawkes process with a constant exogenous intensity of 1.

However, it may not always be possible to obtain an exact value of $\mu(t)$. In such cases, an approximate value can be estimated. To achieve this, we first compute the average intensity function $\lambda_m(t)$ over a span of n days. Each trading day i has the intensity function denoted by $\lambda_i(t)$, and we have a total of n trading days. Specifically, we obtain $\lambda_m(t)$ as the average over all $\lambda_i(t)$: $\lambda_m(t) = \sum_{i=1}^{n} \lambda_i(t)/n$. In practice, we estimate the empirical $\lambda_i(t)$ by normalizing event counts every 100 milliseconds. Assuming that the true exogenous intensity function $\mu(t)$ follows $\lambda_m(t)$, we represent $\mu(t)$ as

$$\mu(t) = \nu(t)\lambda_m(t), \tag{13}$$

where $\nu(t)$ represents the relative exogenous intensity function.

Finally, we define $\xi_i = \int_0^{t_i} \lambda_m(t)\mathrm{d}t$ which serves as timestamps for a Hawkes process M_t with exogenous intensity $\nu(t)$. The resulting process M_t represents the detrended Hawkes process. In general, $\nu(t)$ can be a constant function. However, Filimonov and Sornette (2015) mention that there may be trading days with significant intraday events, making $\nu(t)$ deviate significantly from a constant function. Therefore, we model $\nu(t)$ as a piece-wise constant function to account for this effect.

In the next section, we estimate the piece-wise exogenous function and kernel function of the Hawkes process M_t, which corresponds to the detrended time series. We would like to emphasize that we always detrend with

a one-day background intensity function, but not for other window lengths, although we consider different sub-samples of the same detrended time series in subsequent sections.

4. Model selection

In this section, we perform model selection to determine the appropriate parametric form of the Hawkes process in our empirical analysis.

We assume that the exogenous intensity function $\mu(t)$ is piece-wise constant, where each segment has the same length. This approach is similar to the methodology employed by Wheatley, Wehrli, and Sornette (2019) and Wehrli, Wheatley, and Sornette (2021). Under this assumption, we select the most appropriate parametric form of the kernel function $\phi(t)$ among several candidates, and the optimal number of segments of the exogenous intensity function $\mu(t)$, based on the data of all 108 stocks. In Section 4.1, we first adopt a non-parametric method to estimate the kernel function as the ground truth, and then compare different parametric forms of the kernel function via residual analysis and BIC. In Section 4.2, we select the optimal number of segments s^* of the exogenous intensity function $\mu(t)$ across various window lengths by comparing their respective BIC values.

4.1. Kernel selection

We begin by applying the non-parametric method proposed in (Bacry, Dayri, and Muzy 2012) to estimate the kernel function, which serves as our ground truth. We mainly rely on the auto-covariance function of the Hawkes process, denoted by $\nu_\tau^{(h)}$, where we take h to be the scale and τ to be the lag. Specifically, we define $\nu_\tau^{(h)}$ as follows:

$$\nu_\tau^{(h)} = \frac{1}{h}\mathrm{Cov}\left(N_{t+h} - N_t, N_{t+h+\tau} - N_{t+\tau}\right).$$

In practice, we need to discretize the auto-covariance function $\nu_\tau^{(h)}$ with respect to τ. We define the discretization step size of τ as Δ, and the range of τ as $[0, \tau_{\max}]$.

We apply the non-parametric method to the mid-price processes of all the stocks in our dataset. Following Bacry, Dayri, and Muzy (2012), we set the hyper-parameters $\tau_{\max} = 100$ and $\Delta = h = 0.1$. However, unlike Bacry, Dayri, and Muzy (2012), the assumption of constant exogenous intensity within a trading day is invalid in our dataset (Wehrli, Wheatley, and Sornette 2021). Therefore, we divide each trading day (after detrending) into four equal-sized segments,[10] and assume that the exogenous intensity is constant within each segment. We estimate the auto-covariance function for each segment, compute their average over segments, and then apply the non-parametric method to obtain the kernel function for a single stock. Finally, we use least square estimation to fit the kernel to a power-law function. Appendix 2 gives the specific non-parametric estimation algorithm.

Figure 3 shows the estimated Hawkes kernel function for Ping An Bank, which follows a power-law decay (Bacry, Dayri, and Muzy 2012). We select the fitting interval for the estimated kernel as $[0.5, 100]$ to focus on fitting the tail. We also find this power-law decay phenomenon in all stocks in our dataset as presented in Figure A13 in the Appendix.

Bacry, Dayri, and Muzy (2012) show that non-parametric estimation can accurately reflect the general shape of the kernel function. However, the values and decay rates of the estimated function are greatly influenced by the selection of hyper-parameters, namely Δ and $\tau_{\max}$.[11] Moreover, to estimate the kernel function with greater precision, it is important to have enough samples with consistent exogenous intensity and an identical kernel function. In practical scenarios, these conditions are often unmet. As a result, non-parametric methods can struggle to accurately estimate the endogeneity. To address this issue, we adopt parametric methods and consider several parametric forms of the kernel function in our model selection process. This approach allows us to make individual estimates for each sample.

To apply each parametric form of the kernel function in Section 2.1, we assume that the exogenous intensity function $\mu(t)$ is piece-wise constant over four equally long intervals, which is consistent with the previous non-parametric estimation. Then, we estimate the corresponding parameters by maximum likelihood estimation, as discussed in Section 2.3.

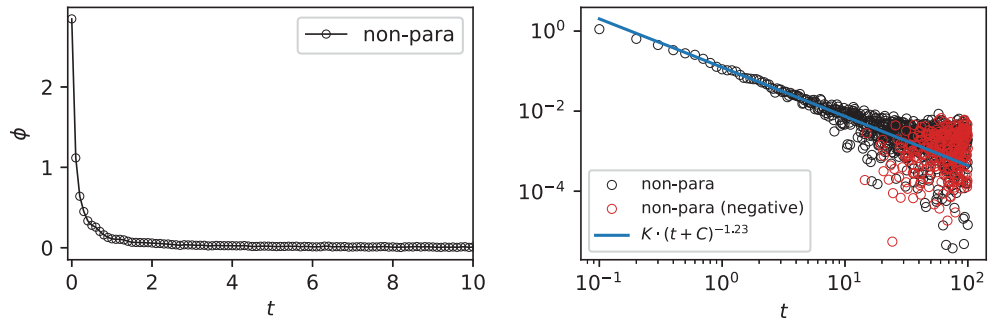
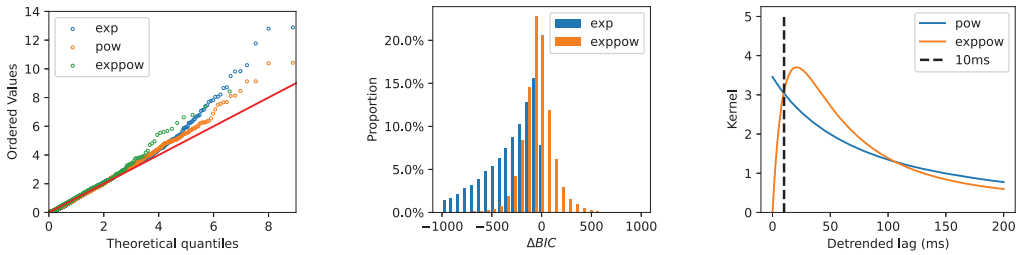


Figure 3. The non-parametric estimate of kernel function for Ping An Bank. The hyper-parameters $\tau_{\max}$, Δ and h are specifically configured as 100, 0.1 and 0.1 respectively. The left panel shows the non-parametric estimate, while the right panel shows the Log-Log plot of the left panel. The black hollows represent the points with positive non-parametric estimates and the red hollows represent the points with negative non-parametric estimates in absolute value. The blue line represents the OLS fit of $\phi(t) = K(t+c)^{-p}$ on [0.5, 100], which is performed at the original (not log-log) scale, and the optimal value of p is approximately 1.23.



(a) QQ-plot of residuals (b) Histogram of ΔBIC (c) Kernel function

Figure 4. Goodness-of-fit analysis of three parametric kernel functions. Each estimation is made over one trading day. Subfigure (a) shows quantile plots of different sequences of $\{\xi_i\}$ against the exponential distribution with rate 1 for exponential law, power-law, and exponential power-law kernels, using Ping An Bank's data on Jan. 2nd, 2020. Subfigure (b) shows the histogram of the ΔBIC's for all stocks over all trading days. Subfigure (c) shows the estimated power-law kernel and exponential power-law kernel for Ping An Bank on Jan. 2nd, 2020. (a) QQ-plot of residuals. (b) Histogram of ΔBIC and (c) Kernel function.

Next, we compare the BIC values of the three kernel functions and perform a residual analysis. The results are presented in Figure 4(a), which indicates that the residuals associated with the power-law kernel exhibit the highest similarity with the residual distribution under the null hypothesis (i.e. the exponential distribution with rate 1).

Across all 108 stocks and all trading days, we calculate the optimal log-likelihood using three kernel functions. Then we subtract the BIC associated with the power-law kernel from the BIC corresponding to the other two kernels, i.e. $\Delta\text{BIC} = \text{BIC}_{\text{pow}} - \text{BIC}$. This ΔBIC serves as a comparative measure to determine whether a certain kernel function is more suitable than the power-law kernel. Figure 4(b) shows the histogram of ΔBIC. For virtually all samples, the BIC corresponding to the power law-kernel is higher than that of the exponential kernel.

However, it is difficult to distinguish power-law kernels from exponential power-law kernels definitively. This challenge may arise from the time resolution of our data being 10 milliseconds, a scale inefficient to capture reactions to market events that occur faster, given the prevalence of high-frequency trading. Figure 4(c) provides an example of the power-law kernel and the exponential power-law kernel estimated in the same process. The two kernel functions are almost the same at the tail, and the first 10 milliseconds show the most difference. In other words, we cannot ascertain the presence of a peak in the kernel function.

To summarize, based on the qualitative observations in Figures 3 and A13, the complexity of the model, and the slightly better performance of the power-law kernel on both the residual and BIC analyzes, we choose the

power-law kernel as our kernel function in subsequent analysis. In fact, we have also performed our analysis using the exponential power-law kernel and found that the empirical results remain almost identical.

4.2. Segments selection

In Section 4.1, we make the assumption that the exogenous intensity is piece-wise constant on four segments. To assess the influence of this choice, we identify the optimal number of segments based on the BIC, as suggested by Wheatley, Wehrli, and Sornette (2019). We find that 3 segments are optimal on average for one trading day.

We begin by introducing some notations. We use T, N_t, and s to denote the window length, the Hawkes process, and the number of segments, respectively. We use n to denote the total number of stock–day, which is approximately 50,000. To investigate the optimal number of segments for various window lengths, we perform analysis for window lengths of 30 minutes, 1 hour, 2 hours, and 4 hours. For the Hawkes process N_t in a given window $[0, T]$, we divide the time window evenly into $s > 0$ segments, namely $\{[iT/s, (i+1)T/s] : i = 0, 1, \ldots, s-1\}$, and assume that the exogenous intensity is a constant and the kernel function is power-law within each segment. To find the optimal s, we consider the BIC, as defined in (9), where the number of parameters n_p is $s + 3$ within this model.

We divide the sample for each trading day into q sub-processes of length T. As discussed in Section 3.2, we detrend each sub-process using its corresponding background intensity. As a result, we have $q \times n$ detrended sub-processes. For the kth sub-process and $s \in \{1, 2, 3, \ldots, 10\}$, we calculate the BIC value corresponding to the maximum likelihood estimation, which is denoted by $\mathrm{BIC}_{T,k,s}$ for $T \in \{30\mathrm{min}, 1\mathrm{h}, 2\mathrm{h}, 4\mathrm{h}\}$ and $1 \leqslant k \leqslant q \times n$. Using the BIC value as a measure of model fit, we define the optimal number of segments corresponding to each sub-process as $s^*_{T,k}$:

$$s^*_{T,k} = \mathrm{argmin}_s \mathrm{BIC}_{T,k,s}. \tag{14}$$

Because the optimal number of segments may vary depending on the specific characteristics of each sub-process, we consider the difference in BIC values ($\Delta\mathrm{BIC}_{T,k,s}$) between the optimal number of segments for each sub-process ($s^*_{T,k}$) and other values of s:

$$\Delta\mathrm{BIC}_{T,k,s} = \mathrm{BIC}_{T,k,s^*_{T,k}} - \mathrm{BIC}_{T,k,s}. \tag{15}$$

For a comprehensive assessment of the impacts of varying segment numbers, we derive an aggregate estimate by calculating the average $\Delta\mathrm{BIC}_{T,s}$ as:

$$\text{Average } \Delta\mathrm{BIC}_{T,s} = \frac{1}{qn} \sum_{k=1}^{qn} \Delta\mathrm{BIC}_{T,k,s}. \tag{16}$$

Figure 5 summarizes the average ΔBIC and the optimal values of s. For the half-hour window and one-hour window, the optimal value of s is 1 based on the average ΔBIC. According to the histogram in Figure 5, for a significant portion of processes with a window length of 4 hours, it is enough to divide into four segments for estimation. This is the reason we use a piece-wise exogenous intensity function with four segments in Section 4.1. Furthermore, the optimal number of segments increases linearly with respect to the window length. For a window length of 4 hours (i.e. the entire trading day), the optimal value for s is 3. Therefore, we fix the number of segments to be 3 in subsequent analysis.

5. Empirical analysis

In this section, we first compare the classical branching ratio (6) with the internal branching ratio (11) that measures the endogeneity of the Hawkes process within a window of finite length.

Based on the internal branching ratio we evaluate the endogeneity of the mid-price process for a single stock, Ping An Bank, under the Hawkes model with a power-law kernel and a piece-wise constant exogenous intensity function. We also discuss the intraday variations of the endogeneity of Ping An Bank.

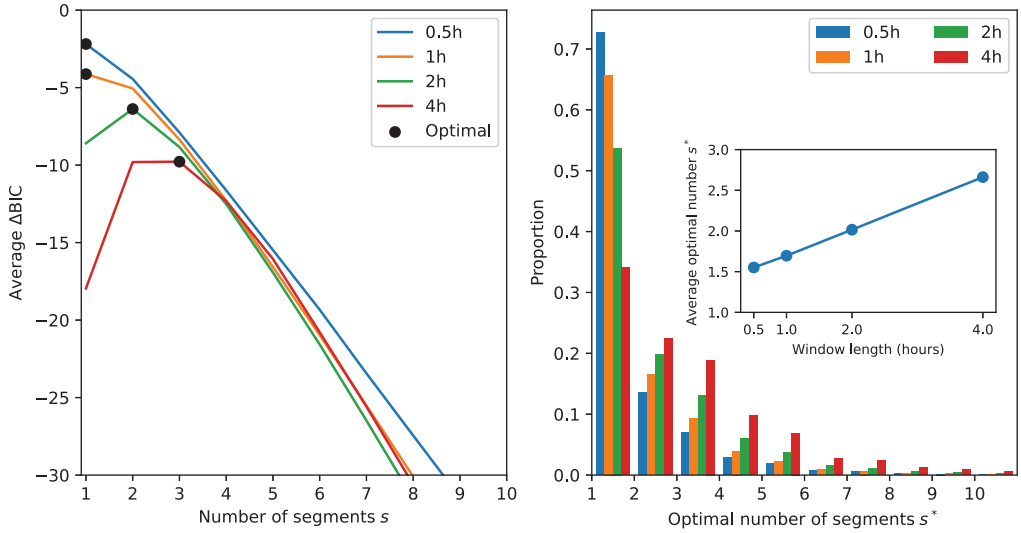


Figure 5. Average ΔBIC and the histogram of the optimal s in percentage terms. For a specific window length T, the timestamp series used for each estimation is a sub-process extracted from Ping An Bank's daily mid-price change process from 2019 to 2020 with a window length of T. See the text for the specific method of obtaining these sub-processes. The left panel shows the average ΔBIC with respect to s, corresponding to distinct window lengths T. The black solid points denote the highest average ΔBIC for each window length T. The right panel shows the histogram of the optimal s (s^*) under different window lengths T. The subplot within the right panel shows the average s^* with respect to the window length T.

We then perform a cross-sectional analysis of the estimated internal branching ratio for all stocks using panel regression. The results show that the internal branching ratio, a price endogeneity measure, is closely related to classical measures of market efficiency at high frequencies. In addition, the price endogeneity in the Chinese market is increasing over time.

5.1. Endogeneity measures

In this subsection, we use real market data to compare the internal branching ratio (11) with the classical branching ratio (6). We first use the maximum likelihood to estimate the classical branching ratio η, $\mu(t_i)$'s and $\lambda(t_i)$'s, and then substitute these estimates into (12) to estimate the internal branching ratio.

Figure 6 compares the estimated internal branching ratio $\hat{\eta}^{\text{in}}$ with $\hat{\eta}$. In addition, we also compare $\hat{\eta}^{\text{in}}$ with the finite integral $\int_0^T \hat{\phi}(t)\mathrm{d}t$, which estimates the expectation of the number of new events generated by an event in the window $[0, T]$, where $\hat{\phi}$ is the estimated kernel. All of these three variables are estimated for each trading day of each stock. The findings indicate a strong positive correlation between $\hat{\eta}^{\text{in}}$ and $\int_0^T \hat{\phi}(t)\mathrm{d}t$. Moreover, when $\hat{\eta}$ is small, $\hat{\eta}$ and $\hat{\eta}^{\text{in}}$ are very close. But when $\hat{\eta}$ is large, the distribution of $\hat{\eta}^{\text{in}}$ appears to be no longer related to $\hat{\eta}$. These observations indicate that $\hat{\eta}^{\text{in}}$ is more appropriate than $\hat{\eta}$ for measuring endogeneity within a finite interval using practical data.

In conclusion, $\hat{\eta}^{\text{in}}$ effectively captures the proportion of endogenous events within the finite window $[0, T]$. Hence, in the subsequent analysis, we use the internal branching ratio $\hat{\eta}^{\text{in}}$ rather than the infinite integral $\hat{\eta}$ to measure the price endogeneity.

5.2. Results for a single stock

In this subsection, we estimate the internal branching ratio using data from a single stock, Ping An Bank (stock code: 000001.SZ), and we focus on the intraday patterns of the internal branching ratio.

Initially, for each trading day (4 hours long), we consider internal branching ratios within the rolling windows with a step size of 3 minutes and a length of 30 minutes. Importantly, these windows exclude both the morning

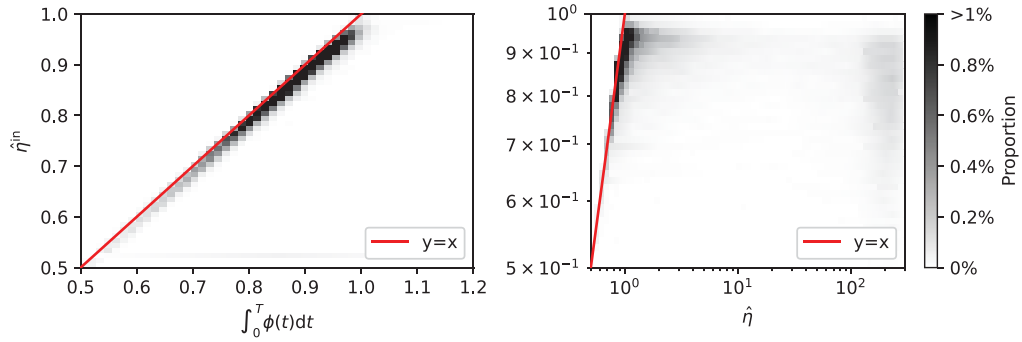


Figure 6. Heatmap of $\hat{\eta}$, $\int_0^T \hat{\phi}(t)\mathrm{d}t$ and $\hat{\eta}^{\text{in}}$. Using the power-law kernel selected in Section 4 and the piece-wise constant exogenous intensity with three segments, we conduct maximum likelihood estimation for each trading day of each stock. Then we use the estimated results to calculate $\hat{\eta}$, $\int_0^T \phi(t)\mathrm{d}t$ and $\hat{\eta}^{\text{in}}$ respectively. The left panel shows a comparison between $\int_0^T \phi(t)\mathrm{d}t$ and $\hat{\eta}^{\text{in}}$, while the right panel shows a comparison between $\hat{\eta}^{\text{in}}$ and $\hat{\eta}$. The red lines in both panels are straight lines aligning with the $y = x$. The shading of each grid indicates the proportion that the sample within that grid to the total.

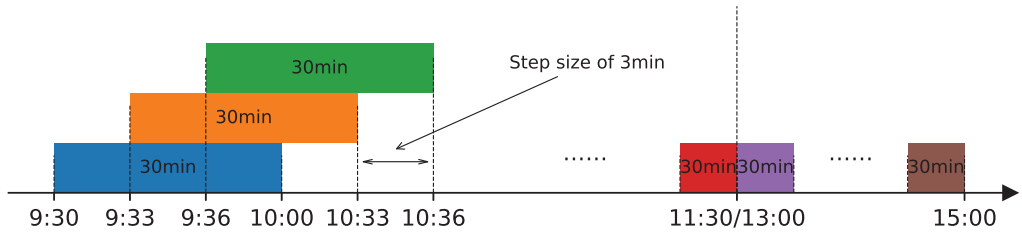


Figure 7. The rolling windows to use for estimation. The window length is 30 minutes, and the step size is 3 minutes.

and afternoon periods. This approach leads to internal branching ratios across 62 windows of length $T = 30\text{min}$. Figure 7 is a schematic diagram of these windows.

To investigate the intraday pattern of the internal branching ratio, we aggregate the data on the windows of the same length, irrespective of their trading dates. After that, we use two different methods to estimate a single internal branching ratio for each time interval: average estimation and aggregated estimation.

For the average estimation method, we first obtain the internal branching ratio on each window by maximum likelihood estimation, and then calculate an average of these internal branching ratios across the windows to yield a single internal branching ratio. Specifically, based on the analysis presented in Section 4.2, we assume a constant exogenous intensity function, $\mu(t) \equiv \mu$, and a power-law kernel function (4) in the Hawkes process model. Let $L_{i,j}(\mu, \eta, p, c)$ denotes the likelihood function of the Hawkes process within the jth window of the ith trading day, where (η, p, c) are the parameters in the power-law kernel function. We obtain the optimal parameters $(\mu, \eta, p, c)_{i,j}$ by maximizing the log-likelihood function, namely

$$\left(\hat{\mu}, \hat{\eta}, \hat{p}, \hat{c}\right)_{i,j} = \text{argmax}_{\mu,\eta,p,c} L_{i,j}(\mu, \eta, p, c). \tag{17}$$

The resulting estimated internal branching ratios form a time series, denoted by $\{\hat{\eta}^{\text{in}}_{i,1}, \hat{\eta}^{\text{in}}_{i,2}, \ldots, \hat{\eta}^{\text{in}}_{i,62}\}$, for the ith trading day. We then calculate the average branching ratios for all trading days by computing the arithmetic mean of the internal branching ratios over the trading days, i.e.

$$\bar{\eta}^{\text{in}}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{\eta}^{\text{in}}_{i,j}, \tag{18}$$

where n represents the total number of trading days.

For the aggregated estimation method, we conduct an aggregated maximum likelihood estimation with a shared branching ratio but varying exogenous intensities along with other kernel function parameters in these windows. Explicitly, assuming that the branching ratio is fixed for n days, with different exogenous intensities for each day, based on a particular kernel function parameter, we calculate the log-likelihood of each trading day with its corresponding exogenous intensity along with other kernel function parameters, and sum them to obtain an aggregated log-likelihood. Optimizing this aggregated log-likelihood also estimates the market's internal branching ratio for a specific time period, that is,

$$\left(\vec{\hat{\mu}}, \hat{\eta}, \vec{\hat{p}}, \vec{\hat{c}}\right)_j = \underset{\vec{\mu},\eta,\vec{p},\vec{c}}{\operatorname{argmax}} \sum_{i=1}^{n} L_{i,j}(\mu_i, \eta, p_i, c_i), \tag{19}$$

where $L_{i,j}$ is the same as in (17), and $\vec{\hat{\mu}}_j$, $\vec{\hat{p}}_j$ and $\vec{\hat{c}}_j$ are the parameter vectors for each day corresponding to the jth time period, respectively. It is actually a $(3n+1)$-dimensional optimization problem. After the estimation, we calculate the internal branching ratio $\tilde{\eta}_j^{\text{in}}$ as a generalization of (11), that is

$$\tilde{\eta}_j^{\text{in}} = \frac{\sum_{i=1}^{n} \sum_{u,v \in S_{i,j}, u<v} \pi_{u,v}^{(i)}}{\sum_{i=1}^{n} N_{i,j}}, \tag{20}$$

where $\pi_{u,v}^{(i)}$ is the probability that the vth event originates from the uth event on the ith day, $S_{i,j}$ is the set of events occurring within the jth period of the ith day, and $N_{i,j}$ is the count of the elements in the set $S_{i,j}$.

Figure 8 shows the results of the internal branching ratios estimated by the two methods mentioned earlier, plotted against the mid-points of time intervals. We observe that the aggregated estimation is always higher than the average estimation. Moreover, the estimated internal branching ratio for each window is generally between 0.65 and 0.9, the average estimate for each period is around 0.79, and the aggregated estimate for each time interval is around 0.81. Interestingly, if we apply the square root of the number of events in the window, namely $\sqrt{N_{i,j}}$, as weights, and calculate the weighted average of the estimated internal branching ratios over days for each period, the resulting values exhibit remarkable similarity to the aggregated likelihood estimation. This suggests that in the aggregated likelihood estimation, the samples with higher internal branching ratios have greater weights, because these samples have more events.

More importantly, both the average and aggregate estimation results show a U-shaped intraday trend, that is, the endogeneity tends to be higher at the beginning of the morning and towards the end of the afternoon. This is similar to the results reported by Wehrli, Wheatley, and Sornette (2021) on the E-mini futures contract, and they claim that this trends arises due to the relatively weak impact of external shocks during these periods. However, it is worth noting that in our dataset, the market intensity function $\lambda(t)_m$ is markedly high at the beginning of the morning and is considerably low towards the close of the afternoon, as illustrated in Figure 2. So this situation does not arise solely because the exogenous intensity is low or high. Instead, we suggest that many numerous market participants concentrate on trading during these two time periods, leading to a significant increase in the endogeneity of the market during these two time periods.

5.3. Cross-sectional analysis

In Section 2.4, we introduce the internal branching ratio as an important measure of price endogeneity, which is related to market efficiency and the rate at which prices absorb external information. In this section, we aim to answer two key questions:

- What is the relationship between the internal branching ratio and market efficiency, and whether it is associated with market efficiency at a particular frequency?
- What factors affect the internal branching ratio?

To answer these two questions, we consider measures of market efficiency, volatility, and other control variables at various frequencies. The literature has proposed several methods to assess market efficiency. Lo and

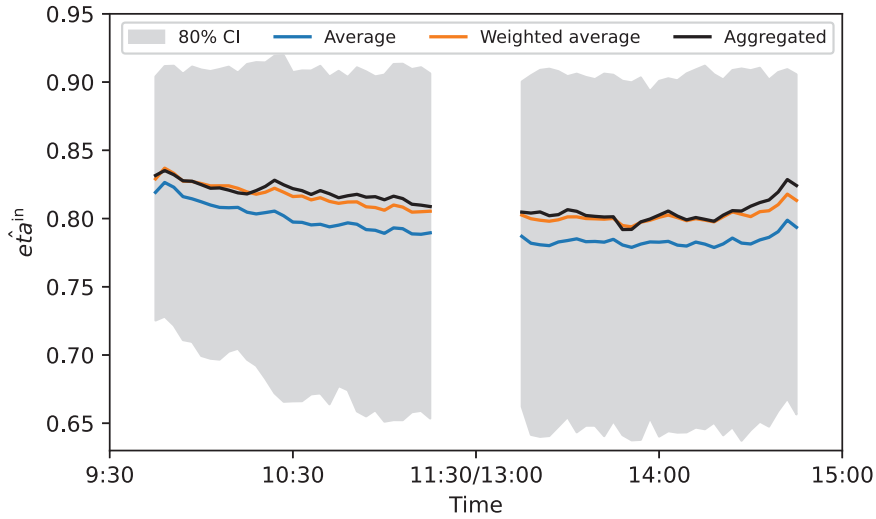


Figure 8. Estimate results of all trading days of Ping An Bank. In the given *j*th window, the figure shows the average, average weighted by the square root of event number, and 80% confidence interval of the sequence of internal branching ratios $\{\hat{\eta}^{\text{in}}_{i,j}\}_{i=1}^{n}$, where *i* represents the *i*th day and *n* represents the total number of trading days. The figure also shows the aggregated estimates for these days. The x-axis corresponds to the midpoint of the respective window. The middle part is left blank because we exclude windows involving both the AM and PM periods.

MacKinlay (1988) use the variance ratio to test whether the stock price follows a random walk; Hou and Moskowitz (2005) propose several price delay measures; Bris, Goetzmann, and Zhu (2007) consider the correlation coefficient between asset returns and lagged market returns. These measures all relate to the second-order moments of returns. Notably, the variance ratio, due to its straightforward form, has found wide application in measuring market efficiency (Borges 2010; Chen, Kelly, and Wu 2020; Chow and Denning 1993; Lo and MacKinlay 1988). We therefore choose this measure to examine how the branching ratio relates to market efficiency.

Table 2 provides an overview of variables included in the regression analysis. Detailed descriptions of these variables are presented in Appendix 3. We mainly focus on the variables related to market efficiency, volatility, and time, and consider other variables as control variables.

These variables of 108 selected stocks for each trading day form a panel dataset. We use panel regression to derive economic insights from the branching ratio. To mitigate the impact of outliers, we apply winsorization to all the variables except the time-related ones. Specifically, we replace the values below the 2.5th percentile and above the 97.5th percentile with the values corresponding to the 2.5th and 97.5th percentiles respectively.

First, we conduct a correlation analysis on several important variables for all trading days for all stocks. Figure 9 illustrates the correlation heatmap of these selected variables. On the one hand, we focus on the correlation between AVR_2's at different frequencies. Specifically, the correlation coefficients between AVR_2^{3s} and AVR_2^{5s}, AVR_2^{20s} as well as AVR_2^{60s} amount to 0.63, 0.02 and -0.08, respectively, which are in descending order. This indicates that the market efficiency is highly correlated at similar frequencies, but is uncorrelated or even slightly negatively correlated at frequencies that diverge significantly. On the other hand, we compare the correlation between the transformed internal branching ratio and AVR_2 at various frequencies. The transformed internal branching ratio is significantly positively correlated with AVR_2^{3s}, a high-frequency measure, and is uncorrelated or negatively correlated with AVR_2's at low frequencies. This observation indicates that the internal branching ratio is more similar in nature to AVR_2^{3s}, i.e. the internal branching ratio reflects price efficiency in terms of absorbing information at high frequencies. Practically, because the estimation of the internal branching ratio is based on the most refined order data, it intuitively encapsulates the high-frequency market characteristics.

To understand what factors affect the branching ratio and study, in particular, whether it is related to classical measures of market efficiency, we perform regression analysis on the panel data. We choose the frequencies of AVR_2^k's, Std^k's and Skewness^k's so that the correlation between independent variables is as low as possible.

Table 2. Variables involved in the correlation and regression analysis. We calculate these variables for each trading day of each stock.

Type	Variable	Description
Dependent Variable	$-\ln(1-\hat{\eta}^{\text{in}})$	The estimated internal branching ratio, after being subjected to a logarithm transformation.
Efficiency	AVR_2^k	The 2-period absolute variance ratio of return at frequency of k.
Volatility	Std^k	The square root of the variance of the return rate at frequency of k.
Time	Days	The number of days counting from Jan. 1st, 2019.
	COVID-19	The 0-1 variable indicating whether the date after the COVID-19, namely after Feb. 3rd, 2020.
Control Variable	$\bar{\mu}^{\text{scaled}}$	The mean of the exogenous intensity function estimated on the detrended data, which is scaled to the magnitude of the corresponding intensity before detrending according to (13).
	Mean Spread	The average of the difference between the best bid price and the best ask price on the order book each time a new order is submitted or canceled.
	Illiquidity Ratio	The ratio of the absolute return for a stock to trading volume over one day. This is also known as Amihud ratio (Amihud 2002).
	Skewness^k	The skewness of the return rate at frequency of k.
	$\log(\text{Market Cap})$	The logarithm of the stock's circulating market capitalization at the close of the trading day.

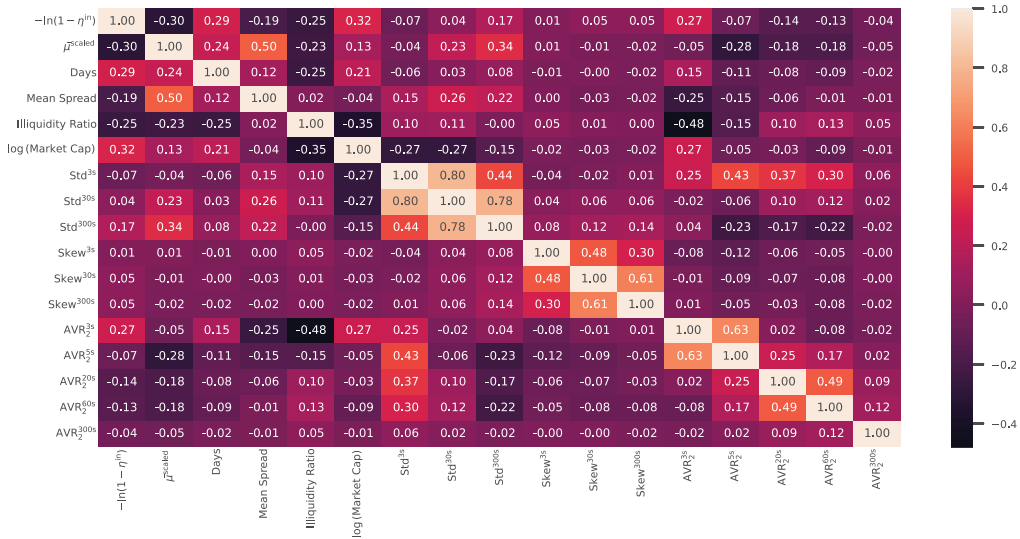


Figure 9. The correlation heatmap of $-\ln(1-\hat{\eta}^{\text{in}})$, $\bar{\mu}^{\text{scaled}}$, days, mean spread, illiquidity ratio, $\log(\text{Market Cap})$, standard deviations and skewnesses at frequencies of 3s, 30s, and 300s, and 2-period absolute variance ratios at frequencies of 3s, 5s, 20s, 60s, and 300s.

Precisely, we choose AVR_2^{3s}, AVR_2^{20s}, AVR_2^{300s}, Std^{3s},Std^{300s}, Skewness^{3s} and Skewness^{300s}. Considering that our panel data contains a total of 108 stocks, we incorporate stock fixed effects in our regression model. Specifically, we use the following regression equation:

$$-\ln(1-\hat{\eta}_{i,t}^{\text{in}}) = \alpha_i + \sum_{j=1}^{13} \beta_j X_{i,t,j} + u_{i,t} \tag{21}$$

Table 3. Panel regression analysis results of the fixed effect model. We calculate the variables given in Table 2 for all trading days for all stocks. In the calculation of the internal branching ratio for each day, the kernel function is a power-law kernel function and the exogenous intensity function is a piece-wise constant function containing 3 segments, as established in Section 4.2. The panel regression analysis involves the transformed internal branching ratio, $-\ln(1-\hat{\eta}^{\text{in}})$, as the dependent variable, while the remaining variables are the treated as independent. *, **, and *** indicate significance at the 90%, 95% and 99% levels respectively.

Variable	Estimated Coefficient
AVR_2^{3s}	0.359**
AVR_2^{20s}	−0.058
AVR_2^{300s}	−0.104***
Std^{3s}	360.970***
Std^{300s}	150.580***
Days	0.001***
COVID-19	0.016
Constant	−6.260***
$\bar{\mu}^{\text{adjusted}}$	−12.188***
Mean Spread	−2.332***
Illiquidity Ratio	−425.750***
Skew^{3s}	−0.022***
Skew^{300s}	−0.001
log(Market Cap)	0.322***
Observations	51,097
R^2	0.5701
F Statistic	5200.366*** (df = 13; 50,976)

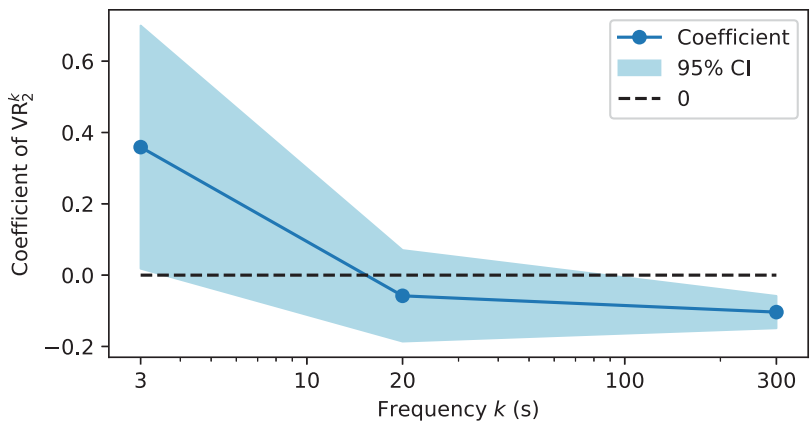


Figure 10. The coefficients of AVR_2 at frequencies of 3s, 20s and 300s in Fixed effect model in Table 3. The light blue shading is the 95% confidence interval of the coefficients.

where $i \in \{1, 2, \ldots, 108\}$ indicates the ith stock, t denotes time, α_i is the intercept of the ith stock, β_j is the slope of the jth variable, $X_{i,t,j}$ is the jth variable of the ith stock at time t, and $u_{i,t}$ is the error term. Table 3 shows the regression results of (21), which we refer to as the fixed effect model.

Market efficiency. In the regression results, the signs and significance of the coefficients regarding AVR_2 vary at different frequencies. Figure 10 shows that the estimated regression coefficients of AVR_2 monotonically decrease with frequency. This observation aligns with intuition: the internal branching ratio of the Hawkes process represents the market characteristics at the highest frequency. This interpretation is reflected by the positive correlation with AVR_2^{3s}, a high-frequency measure, as well as the lack of significant relationship with AVR_2^{20s}, a low-frequency measure. This pattern also corresponds to the findings of the correlation analysis in Figure 9. It is noteworthy that the branching ratio exhibits a significant negative correlation with AVR_2^{300s}.

We conjecture that this phenomenon is primarily driven by the activities of high-frequency traders. When prices deviate from the random walk at high frequencies as measured by the variance ratio statistic, high-frequency traders become more active because they derive profits from these short-term price inefficiencies. Such behaviors reduce pricing errors when measured at low frequencies, making the price more efficient in absorbing external information. This is also consistent with Brogaard, Hendershott, and Riordan's (2014) findings that high-frequency traders gain benefits by predicting short-term price changes, reducing pricing errors, and imposing adverse selection. Our observations support their point of view, but we emphasize that this interpretation is a potential explanation for the observed relationship between the internal branching ratio and measures of price efficiency.[12]

Volatility. Variables Std^{3s} and Std^{300s} measure the volatility of the market at high and low frequencies respectively. The regression results show that both of these two variables have a significant positive coefficient. In fact, the volatility of the market is consistent with the number of changes in the mid-price. Increased market volatility, reflected in more frequent mid-price changes, tends to elevate the degree of market endogeneity, irrespective of whether it occurs at low or high frequencies.

Time. In the Chinese market, we find that the branching ratio is significantly increasing over time, which is consistent with some studies in other markets (Filimonov et al. 2014; Filimonov and Sornette 2012, 2015), while Hardiman and Bouchaud (2014) find it constant. Furthermore, we also find that the COVID-19 crisis has not brought a significant impact on the branching ratio, which is also reported by Yu and Potiron (2022). An increase in the branching ratio is often interpreted as an increase in the proportion of high-frequency strategies that are based on the incoming orders of other traders (Filimonov and Sornette 2012; Wehrli, Wheatley, and Sornette 2021). The increase in the branching ratio in the Chinese market over time is likely due to the increase in the proportion of high-frequency trading.

In general, the internal branching ratio measures the endogeneity of the market, which we find to be associated with market efficiency to some extent. Higher internal branching ratios are associated with lower levels of price efficiency in the low-frequency market. Such a relationship may be due to the fact that high-frequency traders tend to trade when the price at high frequency is inefficient. Furthermore, an increase in market volatility, whether at low or high frequencies, leads to an increase in the internal branching ratio. An important discovery in the Chinese market is that the internal branching ratio is increasing over time, indicating that the market is becoming less efficient at high frequencies. Meanwhile, the analysis indicates that the COVID-19 crisis has not brought significant permanent effects to the market.

6. Robustness checks

We conclude the analysis with a series of robustness checks to understand the potential impact of various modeling and estimation details on the results. First, we examine the influence of the choice of redistribution methods in Section 3 by including a random redistribution method. Second, we investigate how the hyper-parameters $\tau_{\max}$ and Δ in the non-parametric estimation method in Section 4.1 affect the estimation results. Third, we further examine the optimization method in the non-convex maximum likelihood estimation problem (8).

6.1. Redistribution

Because of the time precision, it is possible for multiple mid-price changes to be marked with the same timestamp. As a result, we only have information about the 10ms interval within which the event occurs. To address this issue, we evenly redistribute the events that occur within the same interval. To test the robustness of this operation, we randomly select 5 trading days for each stock and redistribute the events randomly according to the uniform distribution.

Figure 11 gives scatter plots comparing the internal branching ratio using two distinct redistribution methods. The results demonstrate that the estimated internal branching ratios from both redistribution methods are very

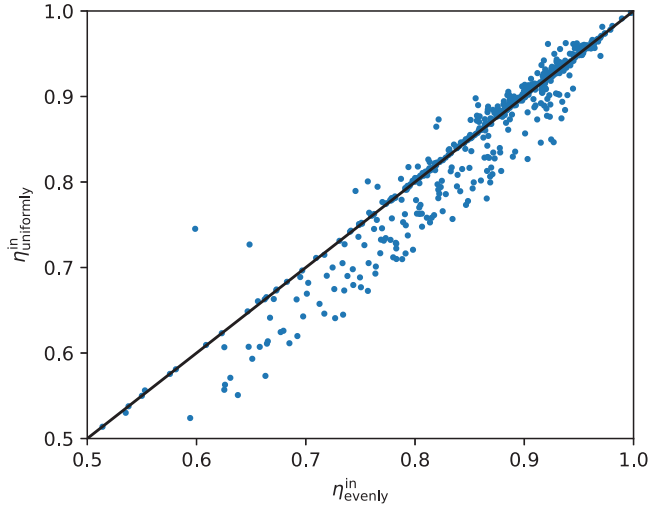


Figure 11. Scatter plots of internal branching ratios obtained by even redistribution and uniform random redistribution.

Table 4. Results of the power law fit $k(t+c)^{p}$, applied to the non-parametric estimated kernel under different hyper-parameters for the mid-price changes of Ping An Bank.

τ_{max}	Δ	k	p	c	τ_{max}	Δ	k	p	c
10.00	0.05	0.1382	1.1317	0.0626	100.00	0.05	0.1332	1.1788	0.0698
	0.07	0.1411	1.1807	0.0763		0.07	0.1368	1.2265	0.0841
	0.09	0.1425	1.2040	0.0816		0.09	0.1386	1.2497	0.0900
	0.10	0.1429	1.2109	0.0824		0.10	0.1390	1.2567	0.0910
	0.12	0.1429	1.2176	0.0808		0.12	0.1393	1.2644	0.0900
	0.20	0.1394	1.2146	0.0560		0.20	0.1365	1.2654	0.0679
	0.50	0.1423	1.3221	0.0000		0.50	0.1373	1.3511	0.0000
50.00	0.05	0.1337	1.1721	0.0687	200.00	0.05	0.1331	1.1821	0.0705
	0.07	0.1371	1.2197	0.0828		0.07	0.1367	1.2297	0.0848
	0.09	0.1389	1.2427	0.0885		0.09	0.1385	1.2530	0.0908
	0.10	0.1393	1.2496	0.0894		0.10	0.1390	1.2600	0.0919
	0.12	0.1395	1.2569	0.0882		0.12	0.1393	1.2679	0.0910
	0.20	0.1365	1.2563	0.0652		0.20	0.1366	1.2699	0.0695
	0.50	0.1379	1.3465	0.0000		0.50	0.1381	1.3685	0.0066

similar. This indicates that the choice of the redistribution method has no significant impact on the estimation of the internal branching ratio $\hat{\eta}^{in}$.

6.2. Non-parametric estimation

Bacry, Dayri, and Muzy (2012) show that the choice of the hyper-parameters Δ and τ_{max} will affect the estimated results. In this part, we perform some experiments to show how the hyper-parameter choices influence our non-parametric estimation results.

Specifically, We estimate the kernel function of the Ping An Bank using the same method in Section 4.1 but with different hyper-parameters. Table 4 displays the hyper-parameters we choose and the corresponding estimate results. The results show that the power exponent p of the estimated kernel function tends to increase with larger discretization step size Δ and window length τ_{max}. Nevertheless, regardless of the parameter choices, we will always get a kernel function with power law decay.

6.3. Maximum likelihood estimation

The non-convex nature of maximum likelihood estimation (MLE) discussed in Section 4.1 for power-law or exponential power-law kernel requires careful optimization. As suggested by numerous existing researches (Bacry et al. 2016; Kirchner and Bercher 2018; Lee and Seo 2017; Yang et al. 2018), the BFGS algorithm performs well in maximizing the log-likelihood function of Hawkes process. In our optimization problem, the variables have bound constraints, so we use the L-BFGS-B algorithm, which extends the BFGS algorithm and is implemented in the Python **SciPy** library. When applying the L-BFGS-B algorithm, we calculate the gradients of the log-likelihood accurately and approximate the Hessian matrix using the algorithm. The stopping criteria are set to achieve a precision of five decimal points.

The existing literature (Filimonov and Sornette 2015; Rizoiu et al. 2017) has also reported local maxima and regions of shallow gradients in the log-likelihood. To mitigate the impact of these issues on our results, we randomly select several different initial values for estimation, which ensures that the result is actually the global maximum rather than a local maximum or in a flat region. Additionally, it is necessary to test the robustness of this procedure. In the following part, we demonstrate that the empirical log-likelihood function has a unique local maximum, which is also global. We also find that the log-likelihood corresponding to the power-law kernel is flat only where p/c is large.

Filimonov and Sornette (2015) define the cost function of the exponential power-law kernel as

$$S(\tau, \epsilon | t_1, t_2, \ldots, t_N) = \min_{\eta, \vec{\mu}} \left(-\ln L(\eta, \vec{\mu}, \tau, \epsilon | t_1, t_2, \ldots, t_N)\right)$$

and provide a surface graph of this function. They indicate that, under the exponential power-law kernel, S can have two local minima. Therefore, when conducting MLE, it is necessary to pay attention to the initial values. We follow their approach to study the property of the cost function. To explain the cost function, we need to first introduce a proposition, which states that the log-likelihood function is convex with respect to some variables. This theorem helps us to observe the shape of the log-likelihood function.

Proposition 6.1: *Under the same notations in Proposition 2.1,* $-\ln L$ *is convex with respect to* $\eta, \mu_0, \ldots, \mu_{p-1}$.

Proof:

$$\begin{aligned}
\ln L &= -\int_0^T \left(\mu(t) + \eta \sum_{t_i < t} h(t - t_i)\right) \mathrm{d}t + \sum_{i=1}^{N_T} \ln \left(\mu(t_i) + \eta \sum_{t_j < t_i} h(t_i - t_j)\right) \\
&= -\int_0^T \mu(t) \mathrm{d}t - \eta \sum_{i=1}^{N_T} \int_{t_i}^T h(t - t_i) \mathrm{d}t + \sum_{i=1}^{N_T} \ln \left(\mu(t_i) + \eta \sum_{t_j < t_i} h(t_i - t_j)\right) \\
&= -\sum_{k=0}^{p-1} \frac{\mu_i T}{p} - \eta H_0 + \sum_{k=0}^{p-1} \sum_{t_i \in \left[\frac{kT}{p}, \frac{(k+1)T}{p}\right)} \ln\left(\mu_k + \eta H_i\right),
\end{aligned}$$

where

$$H_0 = \sum_{i=1}^{N_T} \int_{t_i}^T h(t - t_i) \mathrm{d}t, \quad H_i = \sum_{t_j < t_i, 1 \leqslant j \leqslant N_T} h(t_i - t_j), \quad i = 1, \ldots, N_T.$$

Hence,

$$\frac{\partial^2 \ln L}{\partial \mu_k \partial \mu_l} = \begin{cases} \sum\limits_{t_i \in \left[\frac{kT}{p}, \frac{(k+1)T}{p}\right)} \frac{-1}{\mu_k + \eta H_i}, & k = l, \\ 0, & k \neq l. \end{cases}$$

$$\frac{\partial^2 \ln L}{\partial \eta^2} = \sum_{k=0}^{p-1} \sum_{t_i \in \left[\frac{kT}{p}, \frac{(k+1)T}{p}\right)} \frac{-H_i^2}{\mu_k + \eta H_i}, \quad \frac{\partial^2 \ln L}{\partial \eta \partial \mu_k} = \sum_{t_i \in \left[\frac{kT}{p}, \frac{(k+1)T}{p}\right)} \frac{-H_i}{\mu_k + \eta H_i}.$$

For all $k \in \{0, 1, \ldots, p\}$, we have

$$\begin{vmatrix} \frac{-\partial^2 \ln L}{\partial \eta^2} & \frac{-\partial^2 \ln L}{\partial \eta \partial \mu_0} & \cdots & \frac{-\partial^2 \ln L}{\partial \eta \partial \mu_{k-1}} \\ \frac{-\partial^2 \ln L}{\partial \eta \partial \mu_0} & \frac{-\partial^2 \ln L}{\partial \mu_0^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-\partial^2 \ln L}{\partial \eta \partial \mu_{k-1}} & 0 & \cdots & \frac{-\partial^2 \ln L}{\partial \mu_{k-1}^2} \end{vmatrix}$$

$$= \sum_{l=0}^{k-1} \left(\left(\sum_{t_i \in \left[\frac{lT}{p}, \frac{(l+1)T}{p}\right)} \frac{H_i^2}{\mu_l + \eta H_i} \sum_{t_i \in \left[\frac{lT}{p}, \frac{(l+1)T}{p}\right)} \frac{1}{\mu_l + \eta H_i} - \left(\sum_{t_i \in \left[\frac{lT}{p}, \frac{(l+1)T}{p}\right)} \frac{H_i}{\mu_l + \eta H_i} \right)^2 \right) \cdot \prod_{s \neq l} \sum_{t_i \in \left[\frac{sT}{p}, \frac{(s+1)T}{p}\right)} \frac{1}{\mu_s + \eta H_i} \right)$$

$\geqslant 0$(By Cauchy's inequality).

Thus $-\ln L$ is convex with respect to $\eta, \mu_0, \ldots, \mu_{p-1}$. ■

Thus, if we fix $h(t)$, there exists a unique set of μ's and η such that $\ln L$ reaches its maximum under $h(t)$. This allows us to easily calculate the cost function of kernel $\phi(t)$ numerically, that is:

$$S(h) = \min_{\eta, \vec{\mu}} (-\ln L).$$

As for the power-law kernel, h is determined by two variables c and p. Consequently, we are able to construct a graph of the function S. For different trading days, the function graphs of S are very similar, and we give an example of one trading day in Figure 12. It can be seen that S has a unique local minimum. Notably, S is flat in the upper left part of Figure 12, because in such case p/c is large and η is estimated to be 0, namely that the kernel vanishes. This makes the problem estimate μ only. To avoid this situation, we let $p/c \leqslant 1$ when choosing initial values.

After the above processing and discussions, we conclude that our maximum likelihood estimation is robust. Meanwhile, our estimation results also show that under different initial values, our maximum likelihood estimation always converges to the same point, and the corresponding kernel function at this point is reasonable.

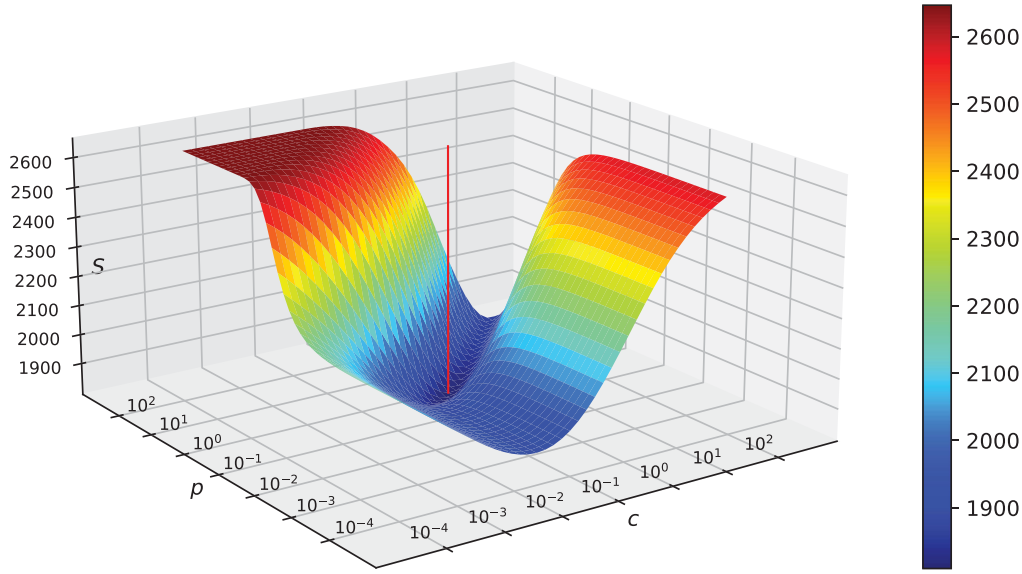


Figure 12. Surface of $S(h)$ calculated by the detrended mid-price changes of Ping An Bank on Feb. 14th, 2019. The vertical red line indicates the global minimum point.

7. Conclusion

In this article, we use a univariate Hawkes process to model the high-frequency mid-price changes of stocks traded on the Shenzhen Stock Exchange in China. Our analysis reveals that the kernel function of price changes in the market follows a power-law decay. We empirically measure the endogeneity within a finite window by the *internal branching ratio*. In addition, our empirical analysis of Ping An Bank shows that the branching ratio has a U-shape intraday pattern, and that high-frequency price changes have significant endogeneity.

We apply our methodology to 108 individual stocks and compare their internal branching ratios cross-sectionally. The results suggest that higher internal branching ratios are associated with lower levels of price efficiency at high frequencies, but higher levels of price efficiency at low frequencies. We conjecture that this is driven by high-frequency trading activities because the market is relatively more inefficient at high frequencies. Furthermore, we find that market volatility will increase the internal branching ratio. Moreover, in the Chinese market, the endogeneity is growing over time and COVID-19 has not brought a significant permanent influence on endogeneity.

In terms of future directions, for the modeling of the kernel function, it is worthwhile to construct a power-law decaying kernel function whose corresponding log-likelihood is convex or has a unique maximum point to make the parameter estimation results more credible. In addition to kernel functions, one can also adopt a more complex model of exogenous intensity to capture the external event shocks. Furthermore, we only consider univariate Hawkes processes in this article. If we treat mid-price increases and decreases as distinct classes of events, we can model them using a multivariate Hawkes process. Similarly, events can also be defined as different orders on the order book. These are important future directions for modeling improvements.

Finally, our study is only a first step towards understanding the microstructure and mechanism of high-frequency information processing in the Chinese market, and relevant research for emerging markets in this direction, in general, is still lacking.

Notes

1. For discussions on this topic, see, for example, Filimonov and Sornette (2012, 2015), Hardiman, Bercot, and Bouchaud (2013), Filimonov et al. (2014), Wheatley, Wehrli, and Sornette (2019), and Wehrli, Wheatley, and Sornette (2021).

2. According to data from World Bank (https://data.worldbank.org), as of 2020, the total value of China's stock market has climbed to a record high of more than USD 12.2 trillion, making it the second-largest in the world after that of the US.
3. Recent empirical evidence shows that the world has started to move from a unipolar to a multipolar financial system in which China plays an increasingly central role (Billio et al. 2022; McKibbin and Fernando 2021).
4. See, for example, Z. Li et al. (2018) and Gao and Ding (2019).
5. Available at: http://www.szsi.cn/cpfw/overseas/market/historical/.
6. CSI 300 is the Chinese equivalent of S&P 500 and it contains 300 securities with large market capitalization and good liquidity. More details are available at https://www.csindex.com.cn/#/indices/family/detail?indexCode = 000300.
7. This includes January 9th, March 14th, and March 15th, 2019.
8. The continuous bidding period for each trading day is from 9:30–11:30 am and 1:00–2:57 pm, which amounts to a total of 3 hours and 57 minutes. To make the morning and afternoon sessions symmetrical for convenience purposes, we expand the afternoon session to 2 hours and simply record the length of a day's trading time as four hours. The last three-minute of the trading day, therefore, do not contain any events. This choice does not bias subsequent estimates because we have detrended the process.
9. Specifically, we redistribute the n events with the same timestamp at t, $t + \Delta/n$, $t + 2\Delta/n$, …, $t + (n-1)\Delta/n$, where $\Delta =$ 10ms is the resolution of our time measurements. Here we adopt a deterministic redistribution approach to improve the reproducibility of our results. In several studies (Filimonov and Sornette 2012, 2015; Hardiman, Bercot, and Bouchaud 2013; Wehrli, Wheatley, and Sornette 2021), the event timestamps are distributed randomly in each interval. In Section 6.1, we verify that our approach leads to very similar estimation results compared to the random redistribution method.
10. According to the selection of the number of segments later in Section 4.2, it is sufficient to divide a day into four segments for estimation.
11. Section 6.2 provides an overview of the impact of hyper-parameter selection on estimation results.
12. To verify that high-frequency traders are driving this phenomenon, one needs transaction-level data with labels of high-frequency traders, which is beyond the scope of this article.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

We thank an anonymous editor and two anonymous reviewers for very helpful comments. Research support from the National Key R&D Program of China (2022YFA1007900), the National Natural Science Foundation of China (12271013), and the Fundamental Research Funds for the Central Universities (Peking University) is gratefully acknowledged.

Notes on contributors

Jingbin Zhuo is a student at Peking University.

Yufan Chen is a student at Peking University.

Bang Zhou is a student at Peking University.

Baiming Lang is a student at Peking University.

Lan Wu is a professor at Peking University.

Ruixun Zhang is an assistant professor at Peking University.

ORCID

Jingbin Zhuo http://orcid.org/0000-0002-6956-0802
Ruixun Zhang http://orcid.org/0000-0002-7670-8393

References

Amihud, Y. 2002. "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects." *Journal of Financial Markets* 5 (1): 31–56. https://doi.org/10.1016/S1386-4181(01)00024-6.

Bacry, E., K. Dayri, and J. F. Muzy. 2012. "Non-Parametric Kernel Estimation for Symmetric Hawkes Processes. Application to High Frequency Financial Data." *The European Physical Journal B* 85 (5): 157. https://doi.org/10.1140/epjb/e2012-21005-8.

Bacry, E., S. Gaïffas, I. Mastromatteo, and J.-F. Muzy. 2016. "Mean-Field Inference of Hawkes Point Processes." *Journal of Physics A: Mathematical and Theoretical* 49 (17): 174006. https://doi.org/10.1088/1751-8113/49/17/174006.

Bacry, E., I. Mastromatteo, and J. F. Muzy. 2015. "Hawkes Processes in Finance." *Market Microstructure and Liquidity* 01 (01): 1550005. https://doi.org/10.1142/S2382626615500057.
Billio, M., A. W. Lo, L. Pelizzon, M. Getmansky Sherman, and A. Zareei. 2022. "Global Realignment in Financial Market Dynamics." SAFE Working Paper.
Borges, M. R. 2010. "Efficient Market Hypothesis in European Stock Markets." *The European Journal of Finance* 16 (7): 711–726. https://doi.org/10.1080/1351847X.2010.495477.
Bowsher, C. G. 2007. "Modelling Security Market Events in Continuous Time: Intensity Based, Multivariate Point Process Models." *Journal of Econometrics* 141 (2): 876–912. https://doi.org/10.1016/j.jeconom.2006.11.007.
Bris, A., W. N. Goetzmann, and N. Zhu. 2007. "Efficiency and the Bear: Short Sales and Markets Around the World." *The Journal of Finance* 62 (3): 1029–1079. https://doi.org/10.1111/j.1540-6261.2007.01230.x.
Brogaard, J., T. Hendershott, and R. Riordan. 2014. "High-frequency Trading and Price Discovery." *Review of Financial Studies* 27 (8): 2267–2306. https://doi.org/10.1093/rfs/hhu032.
Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu. 1995. "A Limited Memory Algorithm for Bound Constrained Optimization." *SIAM Journal on Scientific Computing* 16 (5): 1190–1208. https://doi.org/10.1137/0916069.
Chen, Y., B. Kelly, and W. Wu. 2020. "Sophisticated Investors and Market Efficiency: Evidence From a Natural Experiment." *Journal of Financial Economics* 138 (2): 316–341. https://doi.org/10.1016/j.jfineco.2020.06.004.
Chow, K. V., and K. C. Denning. 1993. "A Simple Multiple Variance Ratio Test." *Journal of Econometrics*58 (3): 385–401. https://doi.org/10.1016/0304-4076(93)90051-6.
Cont, R. 2011. "Statistical Modeling of High-frequency Financial Data." *IEEE Signal Processing Magazine* 28 (5): 16–25. https://doi.org/10.1109/MSP.2011.941548.
Cutler, D. M., J. M. Poterba, and L. H. Summers. 1989. "What Moves Stock Prices?." *The Journal of Portfolio Management* 15 (3): 4–12. https://doi.org/10.3905/jpm.1989.409212.
Fair, R. C. 2002. "Events that Shook the Market." *The Journal of Business* 75 (4): 713–731. https://doi.org/10.1086/341640.
Fama, E. F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383–417. https://doi.org/10.2307/2325486.
Filimonov, V., D. Bicchetti, N. Maystre, and D. Sornette. 2014. "Quantification of the High Level of Endogeneity and of Structural Regime Shifts in Commodity Markets." *Journal of International Money and Finance* 42:174–192. https://doi.org/10.1016/j.jimonfin.2013.08.010.
Filimonov, V., and D. Sornette. 2012. "Quantifying Reflexivity in Financial Markets: Toward a Prediction of Flash Crashes." *Physical Review E* 85 (5): 056108. https://doi.org/10.1103/PhysRevE.85.056108.
Filimonov, V., and D. Sornette. 2015. "Apparent Criticality and Calibration Issues in the Hawkes Self-Excited Point Process Model: Application to High-frequency Financial Data." *Quantitative Finance*15 (8): 1293–1314. https://doi.org/10.1080/14697688.2015.1032544.
Gao, K., and M. Ding. 2019. "Short-Sale Refinancing and Price Adjustment Speed to Bad News: Evidence From a Quasi-Natural Experiment in China." *China Journal of Accounting Research* 12 (4): 379–394. https://doi.org/10.1016/j.cjar.2019.11.001.
Hardiman, S. J., N. Bercot, and J. P. Bouchaud. 2013. "Critical Reflexivity in Financial Markets: A Hawkes Process Analysis." *The European Physical Journal B* 86 (10): 442. https://doi.org/10.1140/epjb/e2013-40107-3.
Hardiman, S. J., and J.-P. Bouchaud. 2014. "Branching-ratio Approximation for the Self-exciting Hawkes Process." *Physical Review E* 90 (6): 062807. https://doi.org/10.1103/PhysRevE.90.062807.
Hawkes, A. G. 1971a. "Point Spectra of Some Mutually Exciting Point Processes." *Journal of the Royal Statistical Society: Series B (Methodological)* 33 (3): 438–443. https://doi.org/10.1111/j.2517-6161.1971.tb01530.x.
Hawkes, A. G. 1971b. "Spectra of Some Self-exciting and Mutually Exciting Point Processes." *Biometrika*58 (1): 83–90. https://doi.org/10.1093/biomet/58.1.83.
Hawkes, A. G. 2018. "Hawkes Processes and Their Applications to Finance: A Review." *Quantitative Finance* 18 (2): 193–198. https://doi.org/10.1080/14697688.2017.1403131.
Hawkes, A. G. 2020. "Hawkes Jump-diffusions and Finance: A Brief History and Review." *The European Journal of Finance* 28 (7): 627–641. https://doi.org/10.1080/1351847X.2020.1755712.
Hawkes, A. G., and D. Oakes. 1974. "A Cluster Process Representation of a Self-Exciting Process." *Journal of Applied Probability* 11 (3): 493–503. https://doi.org/10.2307/3212693.
Helmstetter, A., and D. Sornette. 2002. "Subcritical and Supercritical Regimes in Epidemic Models of Earthquake Aftershocks." *Journal of Geophysical Research: Solid Earth* 107 (B10): ESE 10-1–ESE 10-21. https://doi.org/10.1029/2001JB001580.
Hou, K., and T. J. Moskowitz. 2005. "Market Frictions, Price Delay, and the Cross-Section of Expected Returns." *The Review of Financial Studies* 18 (3): 981–1020. https://doi.org/10.1093/rfs/hhi023.
Jones, C. M., D. Shi, X. Zhang, and X. Zhang. 2020. "Understanding Retail Investors: Evidence from China." Available at SSRN 3628809.
Kirchner, M., and A. Bercher. 2018. "A Nonparametric Estimation Procedure for the Hawkes Process: Comparison with Maximum Likelihood Estimation." *Journal of Statistical Computation and Simulation*88 (6): 1106–1116. https://doi.org/10.1080/00949655.2017.1422126.
Lee, K., and B. K. Seo. 2017. "Modeling Microstructure Price Dynamics with Symmetric Hawkes and Diffusion Model Using Ultra-high-frequency Stock Data." *Journal of Economic Dynamics and Control*79:154–183. https://doi.org/10.1016/j.jedc.2017.04.004.

Li, Z., B. Lin, T. Zhang, and C. Chen. 2018. "Does Short Selling Improve Stock Price Efficiency and Liquidity? Evidence From a Natural Experiment in China." *The European Journal of Finance* 24 (15): 1350–1368. https://doi.org/10.1080/1351847X.2017.1307772.
Li, W., and S. S. Wang. 2010. "Daily Institutional Trades and Stock Price Volatility in a Retail Investor Dominated Emerging Market." *Journal of Financial Markets* 13 (4): 448–474. https://doi.org/10.1016/j.finmar.2010.07.003.
Lo, A. W. 2004. "The Adaptive Markets Hypothesis." *The Journal of Portfolio Management* 30 (5): 15–29. https://doi.org/10.3905/jpm.2004.442611.
Lo, A. W. 2017. *Adaptive Markets: Financial Evolution at the Speed of Thought*. Princeton, NJ: Princeton University Press.
Lo, A. W., and A. C. MacKinlay. 1988. "Stock Market Prices Do Not Follow Random Walks: Evidence From a Simple Specification Test." *Review of Financial Studies* 1 (1): 41–66. https://doi.org/10.1093/rfs/1.1.41.
McKibbin, W., and R. Fernando. 2021. "The Global Macroeconomic Impacts of COVID-19: Seven Scenarios." *Asian Economic Papers* 20 (2): 1–30. https://doi.org/10.1162/asep_a_00796.
Oakes, D. 1975. "The Markovian Self-Exciting Process." *Journal of Applied Probability* 12 (1): 69–77. https://doi.org/10.2307/3212408.
Ogata, Y. 1981. "On Lewis' Simulation Method for Point Processes." *IEEE Transactions on Information Theory* 27 (1): 23–31. https://doi.org/10.1109/TIT.1981.1056305.
Ogata, Y. 1988. "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association* 83 (401): 9–27. https://doi.org/10.1080/01621459.1988.10478560.
Ogata, Y., R. S. Matsuúra, and K. Katsura. 1993. "Fast Likelihood Computation of Epidemic Type Aftershock-sequence Model." *Geophysical Research Letters* 20 (19): 2143–2146. https://doi.org/10.1029/93GL02142.
Ozaki, T. 1979. "Maximum Likelihood Estimation of Hawkes' Self-Exciting Point Processes." *Annals of the Institute of Statistical Mathematics* 31 (1): 145–155. https://doi.org/10.1007/BF02480272.
Rizoiu, M.-A., Y. Lee, S. Mishra, and L. Xie. 2017. "Hawkes Processes for Events in Social Media." In *Frontiers of Multimedia Research*, edited by S.-F. Chang, 191–218. San Rafael, CA: Association for Computing Machinery and Morgan & Claypool.
Saffi, P. A., and K. Sigurdsson. 2011. "Price Efficiency and Short Selling." *Review of Financial Studies*24 (3): 821–852. https://doi.org/10.1093/rfs/hhq124.
Samuelson, P. A. 1965. "Proof that Properly Anticipated Prices Fluctuate Randomly." *Industrial Management Review* 6 (2): 41–49.
Vere-Jones, D. 1970. "Stochastic Models for Earthquake Occurrence." *Journal of the Royal Statistical Society: Series B (Methodological)* 32 (1): 1–45. https://doi.org/10.1111/j.2517-6161.1970.tb00814.x.
Vere-Jones, D., and T. Ozaki. 1982. "Some Examples of Statistical Estimation Applied to Earthquake Data: I. Cyclic Poisson and Self-Exciting Models." *Annals of the Institute of Statistical Mathematics*34 (1): 189–207. https://doi.org/10.1007/BF02481022.
Wehrli, A., S. Wheatley, and D. Sornette. 2021. "Scale-, Time-, and Asset-Dependence of Hawkes Process Estimates on High Frequency Price Changes." *Quantitative Finance* 21 (5): 729–752. https://doi.org/10.1080/14697688.2020.1838602.
Wheatley, S., A. Wehrli, and D. Sornette. 2019. "The Endo–Exo Problem in High Frequency Financial Price Fluctuations and Rejecting Criticality." *Quantitative Finance* 19 (7): 1165–1178. https://doi.org/10.1080/14697688.2018.1550266.
Wu, L., R. Zhang, and Y. Dai. 2022. "Spectral Volume Models: High-Frequency Periodicities in Intraday Trading Activities." Available at SSRN 4230610.
Yang, S. Y., A. Liu, J. Chen, and A. Hawkes. 2018. "Applications of a Multivariate Hawkes Process to Joint Modeling of Sentiment and Market Return Events." *Quantitative Finance* 18 (2): 295–310. https://doi.org/10.1080/14697688.2017.1403156.
Yu, S., and Y. Potiron. 2022. "A Tale of Two Time Scales: Applications in Nonparametric Hawkes Processes with Itô Semimartingale Baseline." Available at SSRN.
Zhang, R., C. Zhao, Y. Chen, L. Wu, Y. Dai, E. Chen, Z. Yao, Y. Zhou, and L. Wu. 2022. "High-Frequency Liquidity in the Chinese Stock Market: Measurements, Patterns, and Determinants." Available at SSRN: https://ssrn.com/abstract = 4191675.
Zhu, C., R. H. Byrd, P. Lu, and J. Nocedal. 1997. "Algorithm 778: L-BFGS-B: Fortran Subroutines for Large-scale Bound-constrained Optimization." *ACM Transactions on Mathematical Software* 23 (4): 550–560. https://doi.org/10.1145/279232.279236.

Appendices

Appendix 1. Power-law Kernel function

In this section, we provide several calculus formulations for the power-law kernel and outline an approximation method for recursively calculating the power-law kernel function.

The power-law kernel function used in this paper is in the form of

$$\phi(t) = \eta p c^p (t + c)^{-1-p}. \tag{A1}$$

Its partial derivatives are

$$\frac{\partial \phi(t)}{\partial \eta} = pc^p(t+c)^{-1-p},$$
$$\frac{\partial \phi(t)}{\partial p} = \eta pc^p(t+c)^{-1-p}\left(1 + p\ln c - p\ln(x+c)\right),$$
$$\frac{\partial \phi(t)}{\partial c} = \eta pc^{p-1}(t+c)^{-2-p}\left(p(t+c) - c(1+p)\right).$$

Let $\int_{t_1}^{t_2} \phi(t)\mathrm{d}t = \Phi(t_1, t_2)$, and $\Phi(t) = \Phi(t, \infty)$, then

$$\Phi(t) = \eta c^p(t+c)^{-p},$$
$$\Phi(t_1, t_2) = \Phi(t_1) - \Phi(t_2),$$

Its partial derivatives are

$$\frac{\partial \Phi(t)}{\partial \eta} = \frac{\Phi(t)}{\eta},$$
$$\frac{\partial \Phi(t)}{\partial p} = \Phi(t)(\ln c - \ln(t+c)),$$
$$\frac{\partial \Phi(t)}{\partial c} = \frac{\eta pc^{p-1}\eta c^p p(t+c)^{-1}}{(t+c)^p} = \Phi(t)\cdot\frac{pt}{c(t+c)}.$$

Define

$$H(s;\eta,p,c) = \frac{\eta pc^p\varphi'\varphi^p}{\Gamma(p+1)e^{c\varphi}}, \tag{A2}$$

where φ is a function of s. We have

$$\int_0^\infty e^{-\varphi t}H(s;\eta,p,c)\mathrm{d}s = \int_0^\infty \frac{\eta pc^p\varphi^p e^{-(c+t)\varphi}}{\Gamma(p+1)}\mathrm{d}\varphi = \phi(t). \tag{A3}$$

Its partial derivatives are

$$\frac{\partial H}{\partial \eta} = \frac{pc^p\varphi'\varphi^p}{\Gamma(p+1)e^{c\varphi}},$$
$$\frac{\partial H}{\partial p} = \frac{\eta c^p\varphi'\varphi^p\left(1 + p\ln(c\phi) - p\psi(p+1)\right)}{\Gamma(p+1)e^{c\varphi}},$$
$$\frac{\partial H}{\partial c} = \frac{\eta p\left(pc^{p-1} - c^p\varphi\right)\varphi'\varphi^p}{\Gamma(p+1)e^{c\varphi}}.$$

From (A3), we give the following discretization algorithm, which is firstly given by Ogata, Matsuúra, and Katsura (1993). Taking $\delta = 1/16$ and considering all i's in $\{0, 1, \ldots, 18\times 16\}$, we calculate the following variables in turn:

$$s_i = -9 + \delta i$$
$$\varphi_i = e^{s_i - e^{-s_i}},$$
$$\log\varphi_i = s_i - e^{-s_i},$$
$$\varphi_i' = e^{s_i - e^{-s_i}}\left(1 + e^{-s_i}\right),$$
$$\log\varphi_i' = s_i - e^{-s_i} + \log\left(1 + e^{-s_i}\right),$$
$$H_i = \frac{\delta\eta pc^p\varphi_i'\varphi_i^p}{\Gamma(p+1)e^{c\varphi_i}}.$$

Hence the discretized counterparts H_i approximate the continuous function H. According to (A3), the multiplication of H_i with the exponential functions $e^{\phi_i t}$ and subsequent summation provide an approximation of $\phi(t)$. Through a recursive calculation of each exponential function $e^{\varphi_i t}$ individually, we establish a method for recursively computing $\phi(t)$. It should be noted that the values of p and c will affect the precision of approximation. In practice, we limit the range of p and c to the interval $[0, 100]$, which is sufficient in our data. Such boundaries ensure that the optimization does not converge to the wrong point due to excessive approximation error.

Appendix 2. Non-parametric estimation

In this section, we outline a non-parametric approach for estimating the kernel function (Bacry, Dayri, and Muzy 2012). Additionally, we present the results of the non-parametric estimation of the kernel function for each individual stock.

First, we introduce some necessary notations. The Laplace transform of a function, f_t, is denoted by

$$\hat{f}_z = \int_{\mathbb{R}} e^{-zt} f_t \mathrm{d}t,$$

and the convolution of two functions A and B is denoted by

$$A \circ B_t = \int_{\mathbb{R}} A_s B_{t-s} \mathrm{d}s.$$

The auto-covariance function of the Hawkes process is denoted by $\nu_\tau^{(h)}$, where we take h as the scale and τ as the lag. Specifically, we define $\nu_\tau^{(h)}$ as:

$$\nu_\tau^{(h)} = \frac{1}{h}\mathrm{Cov}\left(N_{t+h} - N_t, N_{t+h+\tau} - N_{t+\tau}\right).$$

Moreover, we let $\phi_t^{(\circ n)}$ represent the nth auto-convolution of ϕ, and define $\Psi_t = \sum_{n=1}^{\infty} \phi_t^{(\circ n)}$. The following equation helps compute the function $\hat{\Psi}_z$ from the auto-covariance function $\nu^{(h)}$:

$$\left|1 + \hat{\Psi}_z\right|^2 = \frac{\hat{g}_z^{(h)} \bar{\lambda}}{\hat{\nu}_\tau^{(h)}}, \tag{A4}$$

where $g_t^{(h)}$ is defined as:

$$g_t^{(h)} = \begin{cases} 1 - \dfrac{|t|}{h}, & t \in [-h, h], \\ 0, & \text{otherwise}, \end{cases}$$

and $\bar{\lambda} = \frac{\mu}{1-\eta} = \frac{\mathbb{E}[N_t]}{t}$, where $\mathbb{E}[N_t]$ is the expected value of N_t. Substituting this $\hat{\Psi}$ into the equation below yields the Fourier transform of ϕ, namely $\hat{\phi}$:

$$\hat{\phi}_{i\omega} = 1 - e^{-\log|1+\Psi_{i\omega}| + iH\left(\log\left|1+\hat{\Psi}_{i\omega}\right|\right)}, \tag{A5}$$

where H refers to the Hilbert transform and ω is a real number representing frequency. Then we apply the inverse Fourier transform on $\hat{\phi}_{i\omega}$ and obtain ϕ_t.

Next, we introduce the discrete algorithm corresponding to the non-parametric estimation method in Bacry, Dayri, and Muzy (2012):

(1) First we take fixed positive real numbers Δ, h and $\tau_{\max}$ such that

$$K \equiv \frac{\tau_{\max}}{\Delta} \in \mathbb{N},$$

For the timestamp sequence $\{t_i\}$, we calculate its self-covariance sequence $\{\nu_{k\Delta}^{(h)}\}_{|k|\leqslant K}$ at scale h, which is defined as

$$\begin{aligned} \nu_\tau^{(h)} &= \frac{1}{h}\mathrm{Cov}_t\left(N_{t+h} - N_t, N_{t+h+\tau} - N_{t+\tau}\right) \\ &= \frac{1}{h}\left(\mathbb{E}_t\left(\left(N_{t+h} - N_t\right)\left(N_{t+h+\tau} - N_{t+\tau}\right)\right) - (\Lambda h)^2\right), \end{aligned}$$

where $\Lambda = \mathbb{E}_t\left(\lambda_t\right) = \frac{\mathbb{E}_t(\mathrm{d}N_t)}{\mathrm{d}t}$.

In numerical calculation, considering the set of all timestamps as $\mathcal{T}$, with the length of window being T, then the empirical estimates of $v_\tau^{(h)}$'s are obtained as follows:

$$\begin{aligned} v_\tau^{(h)} &= \frac{1}{h\,(T-\tau-h)} \int_0^{T-\tau-h} \left(\int_t^{t+h} \mathrm{d}N_s\right)\left(\int_{t+\tau}^{t+\tau+h} \mathrm{d}N_s\right) \mathrm{d}t - \Lambda^2 h \\ &= \frac{1}{h\,(T-\tau-h)} \int_0^{T-\tau-h} \sum_{\substack{t_1\in\mathcal{T}\\ t\leqslant t_1\leqslant t+h}} \sum_{\substack{t_2\in\mathcal{T}\\ t+\tau\leqslant t_2\leqslant t+\tau+h}} 1\mathrm{d}t - \Lambda^2 h \\ &= \frac{1}{h\,(T-\tau-h)} \sum_{\substack{t_1,t_2\in\mathcal{T}\\ t_2-t_1\in[\tau-h,\tau+h]}} \min\{\tau+h-(t_2-t_1),(t_2-t_1)-(\tau-h)\} \\ &\quad - \Lambda^2 h, \end{aligned}$$

where $\Lambda = \frac{N_T}{T}$.

(2) Then we compute the discrete Fourier transform of $v^{(h)}$:

$$V_k^{(h)} \equiv \sum_{n=-K}^{K} v_{n\Delta}^{(h)} \cdot e^{-\frac{2\pi i}{2K+1}kn},$$

and obtain the sequence

$$V_{-K}^{(h)}, V_{-K+1}^{(h)}, \ldots, V_K^{(h)}.$$

(3) Defining

$$\left|1+\hat{\Psi}_z\right| = \sum_{n=1}^{+\infty} \hat{\phi}_z^n = \frac{\hat{\phi}_z}{1-\hat{\phi}_z},$$

we can prove that

$$\left|1+\hat{\Psi}_{i\omega}\right|^2 = \frac{\hat{v}_{i\omega}^{(h)}}{\bar{\lambda}\hat{g}_{i\omega}^{(h)}},$$

where $\bar{\lambda} = \frac{\mu}{1-\eta} = \frac{\mathbb{E}N_t}{t}$, which represents the average intensity, and

$$g_t^{(h)} = \begin{cases} 1-\dfrac{|t|}{h}, & -h < t < h, \\ 0, & \text{otherwise.} \end{cases}$$

The discrete Fourier transform of $g_t^{(h)}$ is

$$G_k^{(h)} \equiv \sum_{n=-K}^{K} g_{n\Delta}^{(h)} \cdot e^{-\frac{2\pi i}{2K+1}kn}.$$

Thus we can calculate

$$\left|1+\hat{\Psi}_{i\omega}\right| = \sqrt{\frac{V_k^{(h)}}{\bar{\lambda}G_k^{(h)}}},$$

where

$$\omega = \frac{2k\pi}{(2K+1)\,\Delta}, \quad |k| \leqslant K.$$

(4) The Fourier transform of the kernel function ϕ_t, defined as $\hat{\phi}_{i\omega}$, satisfies that

$$\hat{\phi}_{i\omega} = 1 - e^{-\log\left|1+\hat{\Psi}_{i\omega}\right| + iH\left(\log\left|1+\hat{\Psi}_{i\omega}\right|\right)},$$

where H represents the Hilbert transform, and $\omega \in \left\{\frac{2k\pi}{(2K+1)\Delta}\,\middle|\, |k| \leqslant K\right\}$. Hence the Fourier transform of ϕ_t can be calculated.

(5) Finally, we compute the inverse Fourier transform:

$$\phi_{k\Delta} = \frac{1}{2K+1} \sum_{n=-K}^{K} \hat{\phi}_{i\omega_n} \cdot e^{\frac{2\pi i}{2K+1} kn} \left(\omega_n = \frac{2n\pi}{2K+1} \right).$$

The kernel function, ϕ_τ, where $\tau \in \{k\Delta | \, |k| \leqslant K\}$, is obtained. However, it is worth noting that the $\phi(0)$ obtained here is actually $(\phi(0+) + \phi(0-))/2$. The true estimate of $\phi(0+)$ is $2\phi(0)$ because causality dictates that $\phi(0-) = 0$. Therefore, $\phi(0)$ in our estimation results are all multiplied twice to get the true value.

Figure A13 illustrates the non-parametric estimation results of all 108 stocks.

Appendix 3. Variables in regression

In this section, we provide a detailed explanation of the variables discussed in Section 5.3.

A.1 Dependent variable

The dependent variable in the regression is the internal branching ratio. Because the internal branching ratios are bounded in the interval $[0, 1]$ and are often very close to 1, we transform the internal branching ratio using $-\ln(1 - \eta^{\text{in}})$ in our analysis.

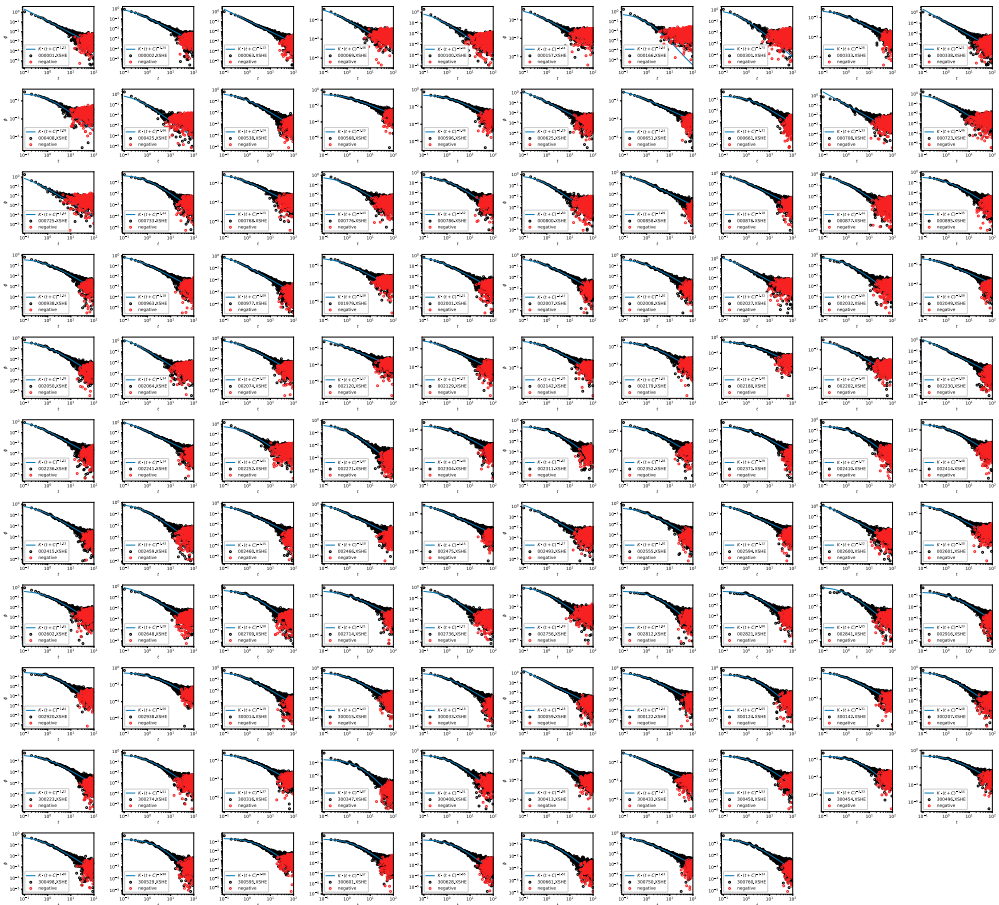


Figure A13. The non-parametric estimate of the kernel function for each of the 108 stocks. The hyper-parameters $\tau_{\max}$, Δ and h are specifically configured as 100, 0.1 and 0.1 respectively. In each sub-panel, the hollow points show the estimated kernel function with their corresponding stock code marked in the legend, and the blue line is the OLS fitting curve of $\phi(t) = K(t + c)^{-p}$ on $t \in [0.5, 100]$, which is the same as in Figure 3.

A.2 Return sequence

We record the return sequence of stock s on day d at frequency k as

$$r_{s,d,i}^{(k)} = \frac{P_{s,d,ik} - P_{s,d,(i-1)k}}{P_{s,d,(i-1)k}},$$

where $i \in \{1, 2, \ldots, T/k\}$, T is the trading hours, namely 4 hours, and $P_{s,d,t}$ is the stock price of stock s on day d at time t.

A.3 Mean spread

We collect the the best bid price and the best ask price sequences in the order book of stock k for day d as $\{b_i\}_{i=1}^N$ and $\{a_i\}_{i=1}^N$ respectively, where i is the ordinal number of the order book update and N is the total number of updates to the order book. Then we calculate the mean spread of stock k on day d as

$$\text{Mean Spread} = \frac{1}{N}\sum_{i=1}^{N}(a_i - b_i).$$

The mean spread serves as an indicator of market liquidity, with a lower value indicating higher market liquidity.

A.4 Variance of return

The variance of return of stock s on day d at frequency q is defined as

$$(\text{Variance of Return})^{(k)} = \text{Var}\left(r_{s,d,i}^{(k)}\right),$$

where the variance is taken over $i = 1, 2, \ldots, T/k$. This variable measures market volatility, with higher values indicating greater volatility. Because of the magnitude of this variable, we use its square root as the independent variable, denoted as the standard deviation:

$$\text{Std}^k = \sqrt{(\text{Variance of Return})^{(k)}}.$$

A.5 Illiquidity ratio

The illiquidity ratio of stock s on day d is defined as

$$\text{Illiquidity Ratio} = \frac{|r_{s,d}|}{V_{s,d}}$$

where $r_{s,d} = P_{s,d}/P_{s,d-1} - 1$, $P_{s,d}$ is the close price of stock s on day d and $V_{s,d}$ is the trading volume of stock s on day d in millions of yuan.

A.6 Absolute variance ratio

The j-period variance ratio of a sequence $\{u_i\}$ is defined as

$$(\text{Variance Ratio})_j = \frac{\text{Var}\left(u_t + u_{t-1} + \cdots + u_{t-j+1}\right)}{j\text{Var}\left(u_t\right)}.$$

In this study, we specifically consider the 2-period variance ratio of the return $\{r_{s,d,i}^{(k)}\}$, that is

$$(\text{Variance Ratio})_2^{(k)} = \frac{\text{Var}\left(r_{s,d,i}^{(k)} + r_{s,d,i-1}^{(k)}\right)}{2\text{Var}\, r_{s,d,i}^{(k)}}.$$

According to the efficient market hypothesis (EMH), when the market is strongly efficient, the price sequence behaves like a random walk, leading to a variance ratio close to 1. Hence the 2-period absolute variance ratio at frequency k, namely

$$\text{AVR}_2^{(k)} = \left|1 - (\text{Variance Ratio})_2^{(k)}\right|,$$

measures the market efficiency to some extent.

A.7 Skewness

The skewness of return of stock s on day d at frequency q is defined as

$$(\text{Skewness of Return})^{(k)} = \frac{\mathbb{E}\left(r_{s,d,i}^{(k)}\right)^3 - \left(\mathbb{E}\, r_{s,d,i}^{(k)}\right)^3}{\left(\text{Var}\, r_{s,d,i}^{(k)}\right)^{\frac{3}{2}}},$$

where the expectation and variance are taken over $i = 1, 2, \ldots, T/k$. This variable provides insight into the asymmetry and shape of the return distribution.