ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
SCHOOL OF
ECONOMIC
SCIENCES

ΤΜΗΜΑ
ΟΙΚΟΝΟΜΙΚΗΣ
ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF
ECONOMICS

Department of Economics

Athens University of Economics and Business

WORKING PAPER no. 06-2024

Time will tell!
Towards the construction of instantaneous indicators of different agent-types

Iordanis Kalaitzoglou, Stelios Arvanitis

July 2024

Πατησίων 76, 104 34 Αθήνα. Τηλ.: 210 8203303-5. E-mail: econ@aueb.gr / www.aueb.gr
76, Patission Street, Athens 104 34 Greece. Tel.: (+30) 210 8203303-5

Πατησίων 76, 104 34 Αθήνα. Τηλ.: 210 8203303-5. E-mail: econ@aueb.gr / www.aueb.gr
76, Patission Street, Athens 104 34 Greece. Tel.: (+30) 210 8203303-5

Time will tell!

Towards the construction of instantaneous indicators of different agent-types

Iordanis Kalaitzoglou [†], Stelios Arvanitis[‡]

[†]Audencia Business School. ikalaitzoglou@audencia.com

[‡]Athens University of Economics and Business, stelios@aueb.gr

July 17, 2024

Abstract

This paper introduces a new instantaneous, intensity-based metric for estimating the proportion of different agent-types at any time interval, coined the (i)ntensity-based (R)elative (P)roportion (iRP). iRP is based on the notion that differences in trading motives are expressed in agent specific arrival rates, which are modelled using conditional hazard functions. This novel framework is applied on identifying the presence of private information and it exhibits empirical and theoretical properties that are superior to existing metrics, even at very short intervals. A variety of other agent-types can be modeled, accordingly, as long as their actions can be mapped into differentials in conditional intensities.

Keywords: Market Microstructure; Agent-types; Hazard Functions.

JEL Codes: C41; C55

1 Introduction

This paper introduces a novel instantaneous estimate for the presence of different agent-types, which can be integrated over any time interval, providing an (i)ntensity-based (R)elative (P)roportion (iRP) metric, suitable for (H)igh (F)requency (T)rading (HFT). The new framework ventures the idea that different trading motives are reflected on tangible actions and, thus, differences on observable variables, aligned with these actions, can reveal the presence of different agents. In particular, in HFT, the time(-ing) of the action is of utmost importance and, thus, differences in the arrival rates should reflect the differential presence of agents. Drawing on the theory of point processes, the conditional intensity (hazard function) is a tool that can describe fully these arrival rates and it is used to capture the presence of these agents. iRP exhibits two major advantages. First, it is interval free and, second, it can be employed to investigate the presence of any agent-type, as long as her actions can be captured by some observable variables. iRP is applied in the context of identifying private information and it exhibits superior theoretical and empirical properties.

Markets, from an asset-pricing point of view, are seen as an information clearing mechanism, where different agent-types interact and resolve their private information. According to the Efficient Market Hypothesis (EMH, [30]), new information is the only reason for permanent price movements, and this process is assumed to be instantaneous and unanimous, since it is the co-ordinated action of rational agents clearing information arbitrage opportunities. However, modern markets, which are mostly dominated by algorithmic trading ([57]), pose a significant challenge to this notion. Algorithms operate at a sub-human attention speeds (<650ms, [43]), rendering the instantaneous price adjustment assumption rather unrealistic. Instead, markets are understood to be semi-strong efficient ([54]), in the sense that private information is incorporated into prices gradually ([12]). This has two major implications.

First, the nature of information that is price resolved changes (e.g., [57]). The EMH advocates that prices are martingales with respect to differential degrees of "fundamental" information about the future cash flows of an asset. That is a lower-frequency, human-like attribute of price discovery that is practically unattainable at the sub-human attention realm of algorithmic trading. Algorithms, with varying degrees of sophistication, extract price related signals from previous trading activity aiming at extracting "trading" information about the fundamental value of the asset, rather than collecting fundamental information. Consequently, the nature of information that affects intraday price discovery is mostly endogenous to trading and requires an assessment of the actions of other market agents, rather than of the magnitude and riskiness of cash flows (e.g., [57]). These signals are mostly related to the arrival rate ([25]) of trading related, observable factors and the direct implication of this is that information and the intensity of these trading factors are inextricably bound in intraday price discovery; highlighting the importance of time/-ing.

Second, the gradual price adjustment, due to the transition from "fundamental" to "trading" information, advocates an implicit classification of actors according to the timing of their access to information. Since information is not price-resolved instantly, it exhibits a "life span", during which different agents might access, interpret and act upon it differently.([57]). Whatever their motivation, they are understood to co-exist in the market and their interactions are the mechanism by which information is price-resolved under a dynamic equilibrium.

Albeit the fact that the agent-type composition is essential for describing this equilibrium, their identification is latent information and cannot be known either ex-post or ex-ante. [5] argues that the identification of individual agent-type strategies is impossible under only public information and, consequently, mean field theory ([45]) cannot address the complexity of this equilibrium, because agent-specific modelling cannot address adequately the emergence properties arising from the interactions of these agents. Consequently, the identification of agent-types cannot be done by designing their individual actions, but rather in a probabilistic way as a sample property.

Toward this end, prior literature suggests using the aggregated characteristics of trading activity and as an observable proxy for the actions of different agent-types, suggesting that they can be used to infer their existence. Considering that at this scale, the timing of events is of utmost importance, the literature focuses, naturally, on trading intensity, complemented by other observable factors. More specifically, since its early stages, the microstructure literature associates the arrival rate of trades ([25]) and volume ([23]; [22]; [21]) with the existence of informed agents. Along the same lines, other agent-types are also identified by the observable characteristics, such as trade sidedness ([62]) associated with asymmetric information, trade initiation runs ([58]) associated with information cascades and herding and trade/order imbalances ([18]; [19]) associated with order flow toxicity. The common denominator of all these metrics is that they are based on the fluctuations in the intensity of a variable and that they derive the probability of the existence of a specific agent-type by the aggregated magnitude over a time-interval.

Albeit intuitive, these metrics suffer from two major issues that might render them practically unusable in HFT; sampling bias and frequency. According to [5] a top-down approach would be more suitable in capturing the emergence properties arising from the interactions of the agents and, thus, interval-based measures are more likely to capture the aggregated properties in manner that is less likely to suffer from noisy signals ([59]). However, there is no clear indication of what an optimal sampling interval might be and empirical evidence (e.g., [18]; [19]; [3]) report a significant sampling bias. They are subject to a trade-off between longer intervals that deal better with noisy signals (e.g., trade initiation, [19]) and shorter ones that are more sensitive to its fluctuations due to mean-reversion properties ([3]). Consequently, higher sampling frequencies that would be more relevant to HFT due to shorter time-intervals, might render interval-based metrics inapt in capturing the finer properties of the data that would reveal the presence of different agent-types, due to noise. In HFT where the timing of possessing information is crucial because it can turn a time priority into an information advantage ([57]); “to be uninformed is to be slow” ([35]). This requires a faster identification of agent-types, ideally as a point rather than as an interval estimate.

This is the primary concern of this study, which suggests a novel, data-driven way to extract the latent information concerning the presence of different agents-types in a probabilistic way as a point estimate. This is done by introducing the instantaneous conditional probability of the arrival of an event type.[1] The arrival rate of different events, such as trades or trading volume, is then linked to the material actions of different agent-types, which should vary according to their motivation for trading. Unlike previous approaches that focus on the 'aggregated properties' of the accumulated actions of each agent-type, this study focuses on the properties of their 'accumulation rates'. The starting point is the presence of the agent-type itself. Each agent-type is assumed to act upon (information) stimuli in a distinct time-invariant way and, consequently, her actions exhibit a distinct time-invariant arrival rate. However, her presence in the market is assumed to be conditional on her access/interpretation of relevant information and thus, her probability of entering the market is conditional. The market as a whole is seen as an infinite mixture (i.e., time-varying probabilities of entering the market) of the (time-invariant) arrival rates of different agent-types and this formulation provides a flexible framework to estimate the probability of an event to be initiated by a particular agent-type. From that, the proportion of this agent-type compared to the total number of trades can be estimated at any time interval.

This approach exhibits several advantages. First, it is interval free, as well as event specific. Instead of estimating the proportion of an agent-type over an interval through the distributional properties of aggregated events, the modelling here uses the instantaneous arrival rates in order to identify the presence of an agent and then it derives the probability of the presence of each agent by the accumulation rate of her actions. This enables its use at any time interval without introducing sampling bias ([3]).

Second, the presence of each agent-type is linked to a precise statistical measure, i.e., the arrival rate/intensity, rather than a specific variable. Previous approaches associate different levels of an observed variable, e.g., trading volume, to a particular agent-type,

[1] The point estimate of the instantaneous conditional probability of the arrival of an event type is the conditional intensity of the point process that describes the arrival rate of this event type. The conditional intensities are a natural measure of arrival rates, which can be interpreted heuristically as "how fast" an event is expected to occur.

such as informed. The approach employed here poses no such restriction. The trading characteristics of an agent-type can be described by any set of observable variables, but the detection of the agent-type is done by how they affect the conditional intensity and not by their magnitude, without assuming any prior link. This provides a data-driven way to extract the latent information about the presence of a multitude of agent-types as long as their fundamental (time-invariant) trading characteristics can be captured by the variation of some observable variables and be mapped into a distinctively shaped hazard function.

Finally, the estimates of the probability of different agent-types is by construction conditional and time variant and thus, it can adapt to evolving market conditions in real time. The modelling approach proposed here satisfies all the properties of a complex system ([44]) by defining the market activity as an infinite mixture of multiple agents that exhibit stationary behaviour (assumed for traceability), who adapt (feedback mechanism) to evolving (evolution) market conditions and thus, their interaction with the market also evolves (becomes non-stationary). Consequently, the modelling here exhibits several desirable properties not present in previous approaches (see Appendix B.1). By assuming stationary behaviour, individual agent-types can be identified, while at the same time, conditioning their interaction with market on market conditions, captures the emergence properties of their collective actions. This duality is done in real time and, therefore, it is relevant to human and algorithmic trading speeds.

The remainder of this paper is organized as follows. Section 2 presents the theoretical framework of how the market is seen as an infinite mixture of time-invariant, agent-specific intensities, as well as how the intensity-based estimator, iRP, can be derived. Section 3 presents an empirical application of iRP on identifying private information. Sections 4 and 5 investigate the empirical and theoretical properties of the new estimator. Further examples and their performance are presented in Appendix B.2. Finally, Section 6 concludes.

2 An Intensity-Based Estimator of Agent-Types

2.1 The Market as a Collection of Different Agents

The market is a collection of $Z = 1, 2, ..., K$ different types of agents that interact and formulate the overall trading activity. Each agent-type is assumed to be driven by a stationary, time-invariant (baseline) motivation for trading, e.g., information, liquidity, technical rule, etc., which determines her (conditional) arrival rate, $\lambda^k\left(t|\mathcal{F}_s\right)$. The arrival rate, $\lambda^k\left(t|\mathcal{F}_s\right)$, can be mathematically described as the conditional intensity of a simple point process.

2.1.1 Agent-Specific Trading as a Simple Point Process

Let $\{T_i\}_{i\in\mathbb{Z}}$ be a simple point process on $[0, \infty)$, defined as a sequence of non-negative random variables on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such as $0 < T_i < T_{i+1} \forall i$.$N(t)$ is the counting process of $\{T_i\}_{i\in\mathbb{Z}}$ defined as $N(t) = \sum_{i\geq 1} 1(T_i < t)$ that counts the events up to time t. Then $\lambda(t|\mathcal{F}_s)$ is defined as the intensity of $N(t)$, given some filtration $\mathcal{F}_s$, if $\mathbb{E}(N(t) - N(s)|\mathcal{F}_s) = \mathbb{E}\left(\int_s^t \lambda(u)\, du \middle| \mathcal{F}_s\right)$, for $0 < s < t$, and fully describes $\{T_i\}_{i\in\mathbb{Z}}$.

The counting function and the arrival rate of each agent-type are defined as follows:

$$N^k(t) = \sum_{i\geq 1} 1(T_i < t)(Z_i = k) \tag{1}$$

$$\mathbb{E}(N^k(t) - N^k(s)|\mathcal{F}_s) = \mathbb{E}\left(\int_s^t \lambda^k(u)\, du \middle| \mathcal{F}_s\right) = \mathbb{E}\left(\int_s^t (p_t^k|\mathcal{F}_s)\lambda_0^k(u)\, du\right) \tag{2}$$

which, defines the arrival of the events of type k as a function of a time-invariant intensity $\lambda_0^k(t)$, which captures the instantaneous probability of a trade of type k to occur, and of a time-variant weighting component $(p_t^k|\mathcal{F}_s) = \mathbb{P}(Z_t = k|\mathcal{F}_s), \sum_{k=1}^K (p_t^k|\mathcal{F}_s) = 1$, which defines the conditional probability of an agent-type k entering the market at a stationary rate of $\lambda_0^k(t)$. The rationale behind this formulation is traceability.

The existence of the different agent-types is known only theoretically and not in the limit when $\Delta t \to 0$. Whether an event is initiated by a specific agent-type, k, cannot be known with certainty and, therefore, its conditional probability is defined as $(p_t^k|\mathcal{F}_s)$. However, in order to trace their existence, their trading characteristics should be able to be mapped into a statistical measure. Following recent literature ([38]; [46]) the trading characteristics of a particular trading behaviour, as they are manifested by an underlying trading motivation, are matched with a distinct shape of the intensity of the relevant point process.

The assumption here is that the underlying characteristics that motivate trading (learning pattern, speed of adjustment, etc.) are time invariant and, thus, stationary. For example, previous literature (eg., [38]) suggests that uninformed traders arrive randomly in the market and, therefore, their arrival rate should be time invariant. This could be mapped into a constant intensity, $\lambda_0^k(t)$. In contrast, technical traders are assumed to extract, analyse and use information in a time-invariant manner. Consequently, their intensity of trading, captured by $\lambda_0^{\text{technical}}(t)$, should not change over time either. This does not mean that the intensity of technical trading is the same (flat) over time. It is a function of time. However, if their trading characteristics, e.g., their learning models or learning speed, do not change, it is the same function of time. This stationarity condition is imposed on all agent-types. However, the conditions that might instigate their trading might change over time, in a way that is conditional on past information. This would result in a time-varying mixture of different agent-types, which is captured by the weighting component, $(p_t^k|\mathcal{F}_s)$.

Collectively, the market can be seen as a time-varying mixture, $(p_t^k|\mathcal{F}_s)$, of different agent-types, $\lambda_0^k(t)$. Unlike the intensities or the weighting components of each agent-type that are latent, the intensity of the market, $\lambda(t|\mathcal{F}_s) = \sum_{k=1}^K \lambda^k(t|\mathcal{F}_s)$, and its counting function, $N(t) = \sum_{i\geq 1} 1(T_i < t) = \sum_{k=1}^K (N^k(t) = \sum_{k=1}^K \sum_{i\geq 1} 1(T_i < t)(Z_i = k)$, are observable. They are the sum of the trading activity of each agent k. The expected number of all events in the market is defined as the sum of all events of all agents present in the market:

$$\mathbb{E}(N(t) - N(s)|\mathcal{F}_s) = \mathbb{E}\left(\int_s^t \lambda(u)\,du\middle|\mathcal{F}_s\right) = \sum_{k=1}^K \mathbb{E}\left(\int_s^t \lambda^k(u)\,du\middle|\mathcal{F}_s\right)$$

$$= \sum_{k=1}^K \mathbb{E}\left(\int_s^t (p_t^k|\mathcal{F}_s)\lambda_0^k(u)\,du\right) \tag{3}$$

Eq.(3) implies that all observed events are necessarily the sum of the events instigated by each agent-type. Consequently, their presence can be 'reversed engineered' from observing the market activity, by estimating a fundamental baseline intensity for each agent-type, as well as a weighting component that captures the probability of its presence.

2.1.2 The Market as an Infinite mixture of Simple Point Process

This formulation has some notable merits that will be used later on in order to construct proxies for latent information. First, it suggests a specific statistical measure, $\lambda_0^k(t)$, that the trading behaviour of different agent-types can be mapped onto. Assuming that the underlying properties of a trading behaviour do not change, they should then be expressed in a particular way. Or else their baseline intensity should be (a) time-invariant (function of time). This property makes it traceable in a data-driven way and this will be the fundamental block of the analysis below. Depending on the particular characteristics of each agent-type, this can be translated into a distinctively shaped intensity, $\lambda_0^k(t)$, that is empirically traceable.

Second, it would be restrictive to assume a time-invariant arrival rate for each agent-type, because this would imply that $\lambda(t|\mathcal{F}_s)$ is also time invariant. Instead, it is more natural to consider that the presence of different agents varies according to evolving market conditions. In Eq.(3) this conditionality is captured by the term $(p_t^k|\mathcal{F}_s)$. This is a weighting component of the baseline intensity, $\lambda_0^k(t)$, which leads to a time-variant intensity for each agent-type, $\lambda^k(t|\mathcal{F}_s)$, the sum of which is the intensity of the market, $\lambda(t|\mathcal{F}_s)$. This way, $\lambda(t|\mathcal{F}_s)$ becomes a weighted average intensity of the baseline intensities, driven by a weighting factor that is conditional on all past information, $\mathcal{F}_s$, up to time $s < t$. $\mathcal{F}_t$ can contain information with respect to past realizations of the full market, $\mathcal{F}_t = \mathcal{F}_t^N$, or the intensity of each agent-type, $\mathcal{F}_t \supseteq \mathcal{F}_t^k, \forall t$, or any other observable variable X, $\mathcal{F}_t \supseteq \mathcal{F}_t^X, \forall t$.

Third, a combination of the two enables a data-driven way to extract information about the existence of different agent-types, as well as their probability of entering the market, from the observed aggregated (superposed process) trading activity in the market. More specifically, the contribution of this paper lies in the interpretation and the modelling of the formulation above and the interpretations of the baseline intensity and its weighting component. Eq.(3) enables the identification of the existence of different agent-types, as long as their trading characteristics can be mapped onto a trading behaviour that exhibits a distinctive intensity, fully captured by $\lambda_0^k(t)$. $\lambda_0^k(t)$ can be empirically extracted from data and this would imply the presence of an agent-type with the 'mapped' trading characteristics. However, this cannot be known with certainty at every point in time and Eq.(3) considers that the superposed process $\lambda(t|\mathcal{F}_s)$ is a weighted average of the baseline intensities $\lambda_0^k(t)$, with weightings $(p_t^k|\mathcal{F}_s)$. This makes $(p_t^k|\mathcal{F}_s)$ a measure of the probability of an event to happen now ($\equiv \lambda(t|\mathcal{F}_s)$) is the instantaneous probability of an event to happen now given that it has not happened so far) to be of type k ($\equiv \lambda_0^k(t)$ is the baseline instantaneous probability of an event of type k to happen now given that it has not happened so far). Thus, $(p_t^k|\mathcal{F}_s)$ can be thought of as the conditional probability of the realization of an event of type k. $(p_t^k|\mathcal{F}_s)$ is conditional on observable information and thus, it can be extracted from the data. This way, Eq.(3) can be used in order to extract information about the existence, $\lambda_0^k(t)$, of different types of agents, their conditional probability, $(p_t^k|\mathcal{F}_s)$, as well as their conditional instantaneous probability,$\lambda^k(t|\mathcal{F}_s) = (p_t^k|\mathcal{F}_s)\lambda_0^k(t)$.

2.2 Toward an Empirical Specification

In order to use Eq.(3) empirically, the conditional intensities, $\lambda^k(t|\mathcal{F}_s)$, must be extracted from the observed intensity of the market, $\lambda(t|\mathcal{F}_s)$. The approach preferred here, over other alternatives, estimates the baseline intensities, $\lambda^k(t|\mathcal{F}_s)$ and the weighting functions, $(p_t^k|\mathcal{F}_s)$, in a data-driven way.[2] Following relevant literature (e.g., [46]), the intensity of the market and of each agent-type are extracted from the conditional distribution $f(t|\mathcal{F}_s)$, and survival function, $S(t|\mathcal{F}_s)$, of arrival times as $f(t|\mathcal{F}_s) = \lambda(t|\mathcal{F}_s)S(t|\mathcal{F}_s)$. Following [29], this is done

[2] Since $\lambda^k(t|\mathcal{F}_s)$'s are not directly observable they cannot be directly estimated using a multivariate Hawke's (e.g., [10]; [37]) process. An alternative approach would be a Markov Renewal (e.g., [60]) process, but this would require a pre-determined number of states, as well as a latent classification.

by focusing on durations, $x_i = t_i - t_{i-1}$, i.e., the time between two consecutive events, i and $i-1$, also defining the information set as $\mathcal{F}_s = \mathcal{F}_{i-1}$. The trading activity of the market as a whole is modelled following an ACD specification:

$$x_i = \psi_i \epsilon_i \tag{4}$$

$\psi_i = \mathbb{E}(x_i|\mathcal{F}_{i-1})$ is the expected duration, conditional on past information $\mathcal{F}_{i-1}$. $\epsilon_i = \frac{\chi_i}{\psi_i}$ is the standardized duration, the distribution of which, $f(\epsilon_i|\mathcal{F}_{i-1})$, is used to derive the distribution of χ_i, $f(\chi_i|\mathcal{F}_{i-1}) = f(\epsilon_i|\mathcal{F}_{i-1})\psi^{-1}$, which in turn is used to derive the conditional intensity for the whole market, $\lambda(\chi_i|\mathcal{F}_{i-1}) = {}^{f(\chi_i|\mathcal{F}_{i-1})}/_{S(\chi_i|\mathcal{F}_{i-1})}$. Following Eq.(3), $f(\chi_i|\mathcal{F}_{i-1})$ and $f(\epsilon_i|\mathcal{F}_{i-1})$ are modelled as a mixture of distributions (e.g., [38]; [33]; [16]):

$$f(\chi_i|\mathcal{F}_{i-1}) = \sum_{k=1}^{K} \mathbb{P}_i(Z_i = k|\mathcal{F}_{i-1}) f^k(x) \tag{5}$$

This specification is consistent with Eq.(3) and defines the distribution of durations at market level as a weighted average of the baseline distributions of the different agent-types k, $f^k(x)$. $f^k(x)$'s are assumed to be time invariant due to time-invariant trading characteristics, but the trigger points that instigate their trading activity are affected by market conditions, $\mathcal{F}_i$. This is captured by the weighting functions $\mathbb{P}(Z_i = k|\mathcal{F}_{i-1})$, which act as time-variant estimates of the probability that the next event will be instigated by agent-type k.

In order for $f^k(x)$ to be traceable, they must be identified based on observable information. A convenient way to do this ([15]; [16]; [46]) is to restrict the difference to a set of $q = 1, ..., Q$ scale/shape parameters of a distribution, different values of which would indicate either a different shape or a different (nested) distribution. Eq.(5) can be rewritten as:

$$f(\chi_i|\mathcal{F}_{i-1}; \boldsymbol{\tau}_i) = f(\chi_i|\mathcal{F}_{i-1}; \boldsymbol{\tau}_i(\boldsymbol{W}_i(:), \boldsymbol{\tau}^m) = \sum_{k=1}^{K} L_i^k(\boldsymbol{W}_i) f^k(x, \boldsymbol{\tau}^k) \tag{6}$$

where, for brevity conditionality can be denoted by the index i and not by $(\cdot|\mathcal{F}_i)$,

$\boldsymbol{\tau}_i = (\tau_i^q)_{q=1}^Q$ is a vector of $q = 1, 2, ..., Q$ parameters, dissected into $m = 1, ..., M$ regimes, that determine the shape and/or the scale of the distribution and, consequently of the conditional intensity, which can then be expressed as a function, $\boldsymbol{\tau}_i(:)$, of a $[Q \times M]$ matrix of weighting functions, $\boldsymbol{W}_i(:)$, and a $[Q \times M]$ vector of shape/scale distribution parameters, $\boldsymbol{\tau}^m = \left(\left(\tau_m^Q\right)_{q=1}^Q\right)_{m=1}^M$. Different combinations of shape/scale parameter τ_m^q estimates would lead to a distribution with a particular shape of the conditional intensity $\lambda^k(x)$ that would match the actions of an agent-type k. The probability of this agent-type k to enter the market, $\mathbb{P}_i(Z_i = k|\mathcal{F}_{i-1})$, can be expressed as a function $L_i^k(:)$ of the weighting functions $\boldsymbol{W}_i$.

Differentiation in Eq.(6) can be reduced to estimable parameters τ_m^q, which can then lead to vectors $\boldsymbol{\tau}^k = \left(\tau_{m:k}^{q=1}, ..., \tau_{m:k}^{q=Q}\right)$, with $m : k$ indicating the identification of regimes m that can lead to an intensity $\lambda^k(x)$ that fully describes the baseline arrival rate of agent-type k.

$$\boldsymbol{\tau}_i = \sum_{k=1}^{K} \mathbb{P}(Z_i = k|\mathcal{F}_{i-1})\boldsymbol{\tau}^k = \sum_{k=1}^{K} L_i^k(\boldsymbol{W}_i)\boldsymbol{\tau}^k \tag{7}$$

Each distribution $f^k(x)$ is fully defined by a set of shape/scale parameters, $\boldsymbol{\tau}^k$, which, according to Eq.(3) describe a time-invariant trading pattern for each agent-type k. The agent-types are not observable and they are inferred from the data by modelling $f(x_i|\mathcal{F}_{i-1} : \boldsymbol{\tau}_i)$ as an infinite mixture with smooth transition functions $\boldsymbol{G}_i$ ([66]) that lead to $\boldsymbol{W}_i$ as:

$$W_{m,i}^q = \left(G_{m,i}^q - G_{m,i+1}^q\right), \textit{ with } G_{1,i}^q = 1,\ G_{M+1,i}^q = 0 \textit{ and } \sum_{m=1}^{M} W_{m,i}^q = 1 \tag{8}$$

$\boldsymbol{G}_i = \left(\left(G_{m,i}^q\right)_{q=1}^Q\right)_{m=1}^M$ is a matrix of smooth transition functions across regimes. They are employed to derive $\boldsymbol{\tau}^k$ in a data-driven way. A combination of $\boldsymbol{G}_i$'s is used as a weighting $\boldsymbol{W}_i = \left(\left(W_{m,i}^q\right)_{q=1}^Q\right)_{m=1}^M$ to capture the conditional probability of being in regime m. This estimate is esential in defining $\mathbb{P}(Z_i|\mathcal{F}_{i-1})$, which can then be expressed as a combination $L_i^k(:)$ of $W_{m:k,i}^q$, where $m : k$ is an indicator of regimes in each shape/scale parameter q that exhibit a shape of the distribution that matches the characteristics of type k.

$$\mathbb{P}(Z_i = k|\mathcal{F}_{i-1} = L_i^k(\boldsymbol{W}_i) = \sum_{Q\bigotimes m:k} \prod_{Q\bigotimes M} W^q_{m,i} \tag{9}$$

The term $\prod_{Q\bigotimes M} W^q_{m,i}$ is the cross-multiplication product of $W^q_{m,i}$ that defines the probability of the intersection between shape/scale parameters, q, and regimes, m, in the contingency table with dimensions M^Q. The term $\sum_{Q\bigotimes m:k}(:)$ identifies all the intersections $(Q\bigotimes m : k)$ where the shape/scale parameters result in a distribution that matches the characteristics of agent k. Then it defines the conditional probability of a trade being instigated by an agent of type $\mathbb{P}(Z_i = k|\mathcal{F}_{i-1})$, as the sum of all these intersections, $Q\bigotimes m : k$.

This specification, following Eqq. (5) - (9), considers that all conditionality is summarized into the functions $\boldsymbol{G}_i$. Consequently, the definition of $\boldsymbol{G}_i$ becomes essential in identifying, in a data-driven way, the number of regimes, as well as the shape of distribution in each regime. A convenient way to do that is by defining them as smooth transition functions ([66]):

$$G^q_{m,i} = G^q_{m,i}(\boldsymbol{J}_i : g^q_m, j^q_m) = \left(1 + e^{-g^q_m(\boldsymbol{J}_i - j^q_m)}\right)^{-1} \tag{10}$$

where, in the spirit of [52], $\boldsymbol{J}_i = \left(\left(J^q_{v,i}\right)^Q_{q=1}\right)^V_{v=1}$, $\boldsymbol{J}_i$ is measurable with respect to $\mathcal{F}_{i-1}$, is a vector of $v = 1, 2, ..., V$ threshold variables (might be a different set for each shape/scale parameter τ^q_i), the level of which in combination with the vector of threshold values, j^q_m, determines the allocation of each event into a regime. Then, each shape/scale parameter, τ^q_i, can be defined as a weighted average of the parameters of each regime, τ^q_m, as:

$$\begin{aligned}\tau^q_i = &\underbrace{\left(\overbrace{G^q_{m=1,i}}^{1} - G^q_{m=2,i}\right)}_{W^q_{m=1,i}} \tau^q_{m=1} + \underbrace{\left(G^q_{m=2,i} - G^q_{m=3,i}\right)}_{W^q_{m=2,i}} \tau^q_{m=2} + ...+ \\ &\underbrace{\left(G^q_{m=M-1,i} - G^q_{m=M,i}\right)}_{W^q_{m=M-1,i}} \tau^q_{m=M-1} + \underbrace{\left(G^q_{m=M,i} - \overbrace{G^q_{m=M+1,i}}^{0}\right)}_{W^q_{m=M,i}} \tau^q_{m=M}\end{aligned} \tag{11}$$

Eq.(11) is a specific derivation of Eq.(7) and defines each shape/scale parameter of the distribution of the amrket point process, τ_i^q, as a wieighted average of the shape/scale parameters, τ_m^q, with m regimes. This way, it becomes possible to match the time-invariant characteristics of an agent-type to a specific parameter, τ_u^q, in a data-driven way. However, although the characteristics of an agent-type might not be time-variant, their presence in the market is. Eq.(11) captures this in the weights $W_{m,i}^q$, which make τ_i^q time-variant. Heuristically, this can be interpreted the following way. The market is composed of different agent-types, whose trading characteristics are time invariant (captured by τ_m^q). The probability that a particular agent-type will instigate the following trade, however, is time varying (captured by $W_{m,i}^q$). This makes the market trading activity (captured by τ_i^q) time varying too. All parameters, including the smoothness parameters, g_m^q, are estimated from the data, which can also determine the number of statistically significant regimes, m, through the significance of the threshold values, j_m^q, and thus, the inferred number of agent-types present in the market.

2.3 The (i)ntensity-based (R)elative (P)roportion

Finally, the expected proportion of an agent-type relative to the total number of events, i.e., intensity-baased Relative Proportion (aka iRP), can be defined using Eq.(3):

$$
\begin{aligned}
(iRP_t^k|\mathcal{F}_s) &= \frac{\mathbb{E}\left(N^k(t) - N^k(s)\middle|\mathcal{F}_s\right)}{\mathbb{E}\left(N(t) - N(s)|\mathcal{F}_s\right)} = \frac{\mathbb{E}\left(\int_s^t \lambda^{Z=k}(u)\, du\middle|\mathcal{F}_s\right)}{\sum_{k=1}^K \mathbb{E}\left(\int_s^t \lambda^k(u)\, du\middle|\mathcal{F}_s\right)} \\
&= \frac{\mathbb{E}\left(\int_s^t \left(p_t^{Z=k}\middle|\mathcal{F}_s\right) \lambda_0^{Z=k}(u)\, du\middle|\mathcal{F}_s\right)}{\sum_{k=1}^K \mathbb{E}\left(\int_s^t \left(p_t^k\middle|\mathcal{F}_s\right) \lambda_0^k(u)\, du\middle|\mathcal{F}_s\right)} = \frac{\left(p_t^{Z=k}\middle|\mathcal{F}_s\right) H^{Z=k}(t|\mathcal{F}_s)}{\sum_{k=1}^K \left(p_t^k\middle|\mathcal{F}_s\right) H^k(t|\mathcal{F}_s)}
\end{aligned}
\tag{12}
$$

Eq.(12) expresses the expected proportion of agent-type k, as a proportion of total events in a general form. Then, considering a specific distribution for each agent-type, the integral $\int_s^t \lambda^{Z=k}(u)\, du$ can be estimated using the cumulative hazard function $H^k(t|\mathcal{F}_s)$. The flexibility of the assumed distribution for the superposed (market) process in exhibiting differently shaped hazard functions (matched to specific trading patterns), becomes crucial. Previous literature (e.g., [16]; [38]; [33]; [46]) consider relatively simple positive support distributions, like the Weibull and/or the Burr distributions, which can only generate monotonically increasing or decreasing hazard functions.

However, the generalization proposed here tries to match the trading behaviour of a considerably wider range of agent-types to the shape of the hazard function and, therefore, monotonic hazard functions might be restrictive. Instead, another distribution is ’indicatively’ proposed here, the q-Weibull distribution, for its relative flexibility in generating non-monotonic hazard functions and its link to ”information entropy”. [3] Following Eqq.(4)-(6), the ”q-Weibull” distribution can be defined for $(\chi_i|\mathcal{F}_{i-1}) \sim W_q\left(A_i = \left[\frac{\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}}}{\psi_i}\right], \tau_i^{q=1}, \tau_i^{q=2}\right)$, as:

$$f(\chi_i|\mathcal{F}_{i-1};\boldsymbol{\tau}_i) = (2-\tau_i^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right) \tag{13}$$

where, A_i is the scale parameter, ”q”:= $\tau_i^{q=1}$ is the entropy parameter, $\tau_i^{q=2}$ is a shape parameter and e_q is the q-Exponential distribution ([11]) that collabses to the exponential when $\tau_i^{q=1} = 1$. The q-Weibull can generate non-monotonic hazard functions:

	$0 < \tau^{q=2} < 1$	$\tau^{q=2} \to 1$	$\tau^{q=2} > 1$
$0 < \tau^{q=1} < 1$	Bath-tub	Increasing	Increasing
$\tau^{q=1} \to 1$	Decreasing	Flat	Increasing
$1 < \tau^{q=1} < 2$	Decreasing	Decreasing	Unimodal

[3] The selection of the q-form is based on ”information entropy” ([67]) or the ”information rate” of a data generation process ([63]) or the ”degree of informativeness” of each observation. This is highly relevant in the venture pursued here, which tries to extract latent information (presence of agent-types) from a noisy signal (the superposed process of the market). The ”information rate” is introduced in the ”q”-form distributions by measuring the impact of the ”surprise” through the Box-Cox transformed parameter $(1-$"q"$)$, which captures the degree of extensivity of the stochastic process that generates the data; or else the impact of the informativeness of an observation in changing the moments of the overall distribution. This is crucial in the framework proposed here because it provides a more flexible way to estimate the state probabilities and how they are affected by the realization of events.

Finally, $H^k(t|\mathcal{F}_s)$ over $\Delta t = t - s$, (e.g., Eq.(12)), can be defined as (e.g., [55]):

$$H(t|\mathcal{F}_s) = \begin{cases} \left(2-\tau^{q=1}\right)\left(A_i(t-s)\right)^{\tau^{q=2}} \sum_{j=0}^{\infty} \frac{[((1-\tau^{q=1})A_i]^j}{j+1} & when\, \tau^{q=1} < 1 \\ \left(A_i(t-s)\right)^{\tau^{q=2}} & when\, \tau^{q=1} \rightarrow 1 \\ \frac{2-\tau^{q=1}}{\tau^{q=1}-1} ln\left[1+\left(\tau^{q=1}-1\right)\left(A_i(t-s)\right)^{\tau^{q=2}}\right] & when\, 1 < \tau^{q=1} < 2 \end{cases} \tag{14}$$

3 Empirical Application: iRP and Information

The empirical specification introduced in Section 2.2 is a generalization of the infinite mixture of distributions methodology (e.g,. [38];[46]) towards an explicit linking with a multitude of agent-types, as long as their trading characteristics can be mapped into a distinctive shape of the hazard function of durations. In more detail, Eq.(12) expresses the relative proportion of a specific agent-type as a function of the cumulative hazard function that describes her 'on average' trading characteristics. Consequently, it is essential to map distinctive trading characteristics to distinctively shaped hazard functions. Eqq.(4)-(11) propose an explicit empirical framework that links observable variables, different regimes of which are associated with different values of shape/scale parameters of the distribution of durations. Naturally, the magnitude of these parameters determine the shape of the distribution, as well as the shape of the hazard function. Therefore, the Eqq.(4)-(12) can identify a plethora of different agent-types, as long as their trading behaviour (i) has a material impact on some observable variables (ii) in a manner that is associated with a distinctively shaped hazard function. There is no constraint on the number of observable variables and/or of distinctive shapes of hazard functions and, thus, the modelling here can theoretically identify the existence of any number of agent-types, as well as their structural changes.

This is a significant generalization over previous approaches and it constitutes the major contribution of this study. The empirical framework proposed in Section 2.2 contributes to the literature in the following ways. First, it proposes a data driven way to a) identify the number of different agent-types in the market b) associate their distinctive trading characteristics to a measurable metric, c) estimate their conditional probability to enter the market and d) ultimately develop a metric for their relative proportion (i.e., iRP in the market. Second, Unlike previous approaches, it develops iRP in the limit when $\Delta t \rightarrow 0$ and not over an interval, which is more appropriate for an HFT environment. This is also done without imposing any prior classification or constrain in the number of observable factors and/or agent-types. Section 3.1 illustrates the mechanics of the proposed modelling with the discussion of an example. Finally, last but not least, another major contribution of this study is the derivation of the limit theory of iRP, which is presented in Section 5.

3.1 (i)ntensity-based (P)robability of (IN)formed trading (iPIN)

One of the most well-documented concepts in the literature ([26]; [56]), is the presence of private information. This information refers to changes in the fundamental value of an asset and it is gradually revealed to the market through the actions of traders that have a timing advantage on it ([57]). These agents, called "informed", as opposed to "uninformed", exploit their information benefit and their actions are revealed to the market by their directional trading. This idea, that deviations from a random arrival of buys and sells (order flow) can carry price-relevant information has been employed by [22, 23] which propose a measure of the (P)robability of (IN)formed trading (PIN). The basic notion in PIN is that when there is no information only liquidity traders exist in the market and the direction of their trading should be random with a probability of 50%. In contrast, when there is private information, informed agents align their demand with the direction of the signal and this creates an order imbalance. The PIN interprets the magnitude of these deviations as increased presence of informed agents. Accordingly, the PIN is defined as, $\frac{\alpha\mu}{(\alpha\mu+e)}$, where a is the probability of the existence of private information and μ and e are the arrival rates of informed and uninformed agents, accordingly. This intuitive measure estimates the proportion of informed agents, i.e., $\alpha\mu$, relative to the number of all trades, i.e., $\alpha\mu + e$.

Albeit insightful, PIN has a notable limitation. It is derived from the trade direction, which is usually latent in raw data. To mitigate the issue of noise in trade direction classification algorithms (e.g., [28]), PIN is estimated over a period of time, without a clear selection criteria for the optimal interval length [4]. This makes it a rather slow interval estimate, inapt for using it a higher trading frequencies required in algorithmic trading ([57]). This is further exacerbated by the fact that PIN is a time invariant estimate. Several approaches propose time variant probabilities conditional on either daily trade imbalances (derived from Bayesian inference [51])) or trade durations (e.g., [20])). These approaches account for differential arrival rates of buys and sells on a daily scale, but they still require a trade classification algorithm and the selection of an optimal interval to mitigate the impact of classification bias; thus, it is relatively slow for high frequency trading standards.

Instead, following a different strand of literature (e.g., [33, 38, 46]) Eq.(12) shifts the focus from the aggregated properties of order flow to the aggregation rate trading to estimate the instantaneous probability of a trade to be informed or uninformed, i.e., $\mathbb{P}(Z_{i=\mathrm{inf}}|\mathcal{F}_{i-1})$ and $\mathbb{P}(Z_{i=\mathrm{uninf}}|\mathcal{F}_{i-1}) = 1 - \mathbb{P}(Z_{i=\mathrm{inf}}|\mathcal{F}_{i-1})$, as well as their relative (time- invariant) arrival rates, i.e., $\lambda_0^{Z=\mathrm{inf}}(t)$ and $\lambda_0^{Z=\mathrm{uninf}}(t)$, respectively. Then using Eq.(12) and considering a time-variant arrival rate of uninformed agents, i.e., $(e_t|\mathcal{F}_s)$ with e_0 being their baseline intensity, PIN can be transformed into an intensity-based equivalent, iPIN:

$$
\begin{aligned}
\mathrm{iPIN}_t = \left(\mathrm{iRP}_t^{\mathrm{inf}}\middle|\mathcal{F}_s\right) &= \frac{(\alpha_t|\mathcal{F}_s)\mu}{(\alpha_t|\mathcal{F}_s)\mu + \underbrace{\{1 - (\alpha_t|\mathcal{F}_s\}\, e_0}_{(e_t|\mathcal{F}_s)}} = \frac{\mathbb{E}\left(\int_s^t \lambda^{Z=\mathrm{inf}}(u)\,du\middle|\mathcal{F}_s\right)}{\sum_{k=1}^K \mathbb{E}\left(\int_s^t \lambda^k(u)\,du\middle|\mathcal{F}_s\right)} \\
&= \frac{\mathbb{E}\left(\int_s^t \left(p_t^{Z=\mathrm{inf}}\middle|\mathcal{F}_s\right)\lambda_0^{Z=\mathrm{inf}}(u)\,du\right)}{\sum_{k=1}^K \mathbb{E}\left(\int_s^t \left(p_t^k\middle|\mathcal{F}_s\right)\lambda_0^k(u)\,du\right)} = \frac{\left(p_t^{Z=inf}\middle|\mathcal{F}_s\right) H^{Z=inf}(t|\mathcal{F}_s)}{\sum_{k=1}^K \left(p_t^k\middle|\mathcal{F}_s\right) H^k(t|\mathcal{F}_s)}
\end{aligned}
\tag{15}
$$

which, can then be expressed in terms of the regimes of shape parameters, τ_i^q, as:

[4] ([3]). [22, 23] suggest that a time interval of approximately one month produces a sufficient quantity of data to estimate PIN with relative accuracy; a claim that has been debated in the literature (e.g., [59])

$$
\begin{aligned}
\mathrm{iPIN}_i &= \frac{\mathbb{E}\left(\int_s^t L_i^{Z=\mathrm{inf}}(\boldsymbol{W}_i)\lambda_0^{Z=\mathrm{inf}}\left(u, \boldsymbol{\tau}^{Z=\mathrm{inf}}\right)\,du\right)}{\sum_{k=1}^{K}\mathbb{E}\left(\int_s^t L_i^{k}(\boldsymbol{W}_i)\lambda_0^{k}\left(u, \boldsymbol{\tau}^{k}\right)\,du\right)} = \\
&= \frac{\sum_{Q\bigotimes(m:k=\mathrm{inf})}\left\{\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes(m:k=\mathrm{inf})}(t)\right\}}{\sum_{Q\bigotimes M}\left\{\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes M}(t)\right\}} \\
&= \frac{\sum_{Q\bigotimes\left(m\left(\tau_m^{q=1}\geq 1,\tau_m^{q=2}\leq 1\right):k\right)}\left\{\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes\left(m\left(\tau_m^{q=1}\geq 1,\tau_m^{q=2}\leq 1\right):k\right)}(t)\right\}}{\sum_{Q\bigotimes M}\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes M}(t)} \qquad (16)
\end{aligned}
$$

Eqq. (15)-(16) express the probability of informed trading ($Z = \mathrm{inf}$) in terms of estimable parameters. The parameters in τ^k determine the shape of the hazard function that is then matched to the trading characteristics of different agent-types. The parameters in W_i capture the probability of belonging to different regimes, m, and consequently, the conditional probabilities that a trade is initiated by a particular agent-type. The numerator is the number of trades initiated by informed agents and the denominator is the total number of trades. This is a definition identical to conventional PIN, but Eq.(16) provides an instantaneous estimate of PIN that can be estimated conditionally over any interval.

The last line of Eq.(16) expresses iPIN explicitly when the characteristics of informed agents can be described by more threshold variables and/or shape/scale parameters. A matching shape (decreasing) of the hazard function, would be observed when $\tau^{q=1} \geq 1$ and $\tau^{q=2} \leq 1$. In this formulation, the regimes, m (might be different for each q) of $\tau^{q=1}$ $\tau^{q=2}$ that lead to a decreasing shape of the hazard function, identify informed trading ($k = inf$). Then, according to Eq.(12, the aggregated number of informed agents over time is the sum of all the probabilities, i.e., $\prod_{Q\bigotimes M} W_{m,i}^{q}$, of intersections $Q\bigotimes\left(m\left(\tau_m^{q=1} \geq 1, \tau_m^{q=2} \leq 1\right) : k\right)$ in the contingency table, where $\tau^{q=1} \geq 1$ and $\tau^{q=2} \leq 1$, times the respective cumulative hazard functions of these intersections, i.e., $H^{Q\bigotimes\left(m\left(\tau_m^{q=1}\geq 1,\tau_m^{q=2}\leq 1\right):k\right)}(t)$. This is then compared to the expected number of the trades of all agent-types, defined explicitly in the denominator. A specific version of Eq.(16), as an example, is discussed in Section 4.

3.2 VPIN: A High(er) Frequency PIN and iVPIN

A fundamental difference of iPIN compared to the conventional PIN is that it derives the probability of informed trading not from the aggregated sign of trades, but from two elements of the trading activity; trading frequency and trading volume. This is in line with previous studies that challenge the relevance of the original version of the PIN, and account for the time dimension (e.g., [20]), as well as for the volume dimension (e.g., [18, 24]). Since the beginning the literature recognised that the PIN is a rather noisy measure, primarily due to its reliance on the noisy signal of trade initiation, with different time interval lengths changing its distributional properties ([3]), rather than mitigating the issue. In a series of studies (e.g., [18, 24]), [18] relax both assumptions and suggest a(n), more HFT-friendly, estimate for the probability of informed trading, based on fixed buckets of volume or time, as well as from price changes (rather than trade imbalances). The new measure, named VPIN, is an explicit recognition that the speed of volume accumulation might be more strongly associated with information. Furthermore, aggregated signed volume, which could be thought of as Volume Imbalance, VI_i, in the limit approaches $\alpha\mu$ (informed trades). In parallel, the total aggregated volume accounts for all trading and thus for $\alpha\mu + e$. Then, using a rolling window of length n, VPIN_t can be estimated as $\text{VPIN}_{t_{bucket}} = \frac{\sum_{t_{bucket}-n}^{t_{bucket}} \left| V^B_{t_{bucket}} - V^S_{t_{bucket}} \right|}{\sum_{t_{bucket}-n}^{t_{bucket}} V_{t_{bucket}}}$, with $\frac{\mathbb{E}(\|V^B - V^S\|)}{\mathbb{E}(V^B + V^S)} \to \frac{\alpha\mu}{(\alpha\mu + e)}$, where $\sum_{t_{bucket}-n}^{t_{bucket}} \left| V^B_{t_{bucket}} - V^S_{t_{bucket}} \right|$ is the volume imbalance and $\sum_{t_{bucket}-n}^{t_{bucket}} V_{t_{bucket}}$ is the total volume over the last n buckets preceding bucket-time t_{bucket} .

Extending on this idea, Eq.(12), can be parameterised to account for the arrival rate of volume and associate it with different agent-types. Whereas VPIN approaches informed trading from the perspective of aggregated outcome of signed volume, the approach here focuses on the aggregation process itself. iPIN in Eqq.(15)-(16) is based on the counting function of the trades instigated by different agent-types, which renders it a one dimensional (trades) metric that is inadequate to capture the volume dimension. $N(t)$ counts the total number of trades over an interval, but not the total volume. For this purpose a reformulation of t is required. Previous literature provides various ways that the arrival times t_i can be transformed to account for a higher dimension of additional "marks", but all boil down to the concept of measuring the arrival rate of a "differently" defined event. [29] suggest focusing on the waiting time for a price change of certain magnitude, a term coined "price duration", while [47] focus on the waiting time of a unit magnitude of an associated mark, e.g., the waiting time per unit of volume. Accordingly, $t \rightarrow t^*$, where the counting process $N(t^*) = \sum_{i\geq 1} 1(T_i < t^*)$ counts how many events, described by the additional dimension, occur over a time interval. Define $t^* = t^*_i - t^*_{i-1}$ as the waiting time for a given magnitude of volume ([47]). $N(t^*)$ counts how much volume is traded over a time interval and following Eqq.(2)-(3) it can be traced back to the agent-type who initiated it. $N(t^*)$ can be defined as $N(t^*) = \sum_{i\geq 1} 1(T_i < t*) = \sum_{k=1}^{K} \sum_{i\geq 1} 1(T < t*)(Z_i = k)$. This way, each counting function $N^k(t^*)$ counts the volume traded by each agent-type k. With this formulation, the (volume) trading activity of each agent-type is modelled and its relative proportion to the total trading activity can be computed according to Eq.(12). Accordingly, in a similar fashion to iPIN , the intensity-based VPIN, i.e., iVPIN, can be formulated as:

$$\text{iVPIN}_i = \frac{\sum_{Q\bigotimes\left(m\left(\tau_m^{q=1}\geq 1,\tau_m^{q=2}\leq 1\right):k\right)} \left\{\prod_{Q\bigotimes M} W^q_{m,i} H^{Q\bigotimes\left(m\left(\tau_m^{q=1}\geq 1,\tau_m^{q=2}\leq 1\right):k\right)}(t^*)\right\}}{\sum_{Q\bigotimes M} \prod_{Q\bigotimes M} W^q_{m,i} H^{Q\bigotimes M}(t^*)} \tag{17}$$

The formulation above is different from Eq.(16) in the way the conditional intensities are considered. Instead of the intensities $\lambda^k(t)$ that account for the arrival rate of the transactions instigated by agent-type k, Eq.(17) is based on the intensities $\lambda^k(t^*)$ that account for the arrival rate of volume traded by agent-type k. This way, in Eq.(17), the regimes, m of $\tau^{q=1}$ and of $\tau^{q=2}$ (m could be different for each parameter) that lead to a decreasing shape of the hazard function, identify informed trading ($k = inf$), by considering the arrival rate of their trading measured in terms of volume. Then the numerator defines the aggregated volume traded by informed agents over time as the sum of all the probabilities, i.e., $\prod_{Q\bigotimes M} W^q_{m,i}$, of intersections $Q \bigotimes \left(m\left(\tau_m^{q=1} \geq 1, \tau_m^{q=2} \leq 1\right) : k\right)$ in the contingency table, where $\tau_m^{q=1} \geq 1$

and $\tau_m^{q=2} \leq 1$, times the respective cumulative hazard functions of these intersections, i.e., $H^{Q \otimes \left(m\left(\tau_m^{q=1} \geq 1, \tau_m^{q=2} \leq 1\right):k\right)}(t^*)$. This is then compared to the aggregated total volume that is defined in the denominator. This is an alternative measure to VPIN in continuous time. Again, a specific parameterization can be found in Section 4.

iVPIN exhibits notable advantages over VPIN, on top of the fact that it is not an interval measure. The empirical properties of VPIN depend on the selection ([3]) of the time (e.g., [18, 24]) or volume ([17]) bucket size, as well as on trade classification. iVPIN, instead of selecting an "optimal" time or volume size that could potentially account for the aggregation properties of information, it models exactly this process with the implied intensities of the different agent-types. Consequently, iVPIN, unlike VPIN, derives the presence of private information from the relative speed that volume accumulates, rather than from the direction of accumulated volume and its sampling properties.

4 Empirical Properties of the iRP

The objective of this section is to investigate the empirical (Section 4.2) properties of the iRP measure (Eq.(12)) before investigating its theoretical properties. This is done in the following way. First, Section 4.1 presents an example of an empirical specification for iVPIN and discusses how the relevant probabilities can be linked to specific, estimable parameters. Then, Section 4.2 provides empirical estimates of these parameters and compares the empirical performance of iVPIN to other conventional metrics. Then, the following section (5) discusses the theoretical properties of iVPIN as an example metric of iRP.

4.1 From iRP to iPIN and iVPIN. An Empirical Specification

Drawing on previous literature (e.g., [46, 47]), in order to illustrate how Eqq.(16, 17) can be linked to estimable parameters, the following specification of iVPIN is employed in all the analysis below, as an indicative example:

Table 1: Indicative Empirical Specification Example

$$\chi_i = \psi_i \epsilon_i,\ \psi_i = \mathbb{E}\left(\chi_i | \mathcal{F}_{i-1}, \omega, \beta, \phi, \delta\right) = \omega + \beta\psi_{i-1} + \left(\chi_i - \beta\chi_{i-1}\right) - \left(\tilde{\chi}_i - \phi\tilde{\chi}_{i-1}\right)$$

$$f\left(\chi_i | \mathcal{F}_{i-1}; A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}} / \psi_i\right], \tau^{q=1}, \tau_i^{q=2}\right) = \left(2-\tau^{q=1}\right)\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right)$$

$$\tau_i^{q=2} = \underbrace{\left(\overbrace{G_{m=1,i}^{q=2}}^{1} - G_{m=2,i}^{q=2}\right)}_{W_{m=1,i}^{q=2}} \tau_{m=1}^{q=2} + \underbrace{\left(G_{m=2,i}^{q=2} - \overbrace{G_{m=3,i}^{q=2}}^{0}\right)}_{W_{m=2,i}^{q=2}} \tau_{m=2}^{q=2}$$

$$\text{for} \quad G_{m=2,i}^{q=2} = \left(1 + e^{-g_{m=2}^{q=2}\left(J_i - j_{m=2}^{q=2}\right)}\right)^{-1}, \text{ with } J_i = ti_i = \left(duration_i * K(volume_i)\right)^{-1}$$

where, χ_i being defined as $\chi_i = \Delta t_i$, when a trade-"clock" is employed (relative to PIN), or as as $\chi_i = \Delta t_i^*$, when a volume-"clock" is employed (relative to VPIN). The conditional mean follows a FIACD$(1, \delta, 1)$ (e.g., [41]) specification, where $\tilde{\chi}_i = (1-L)^{\delta}\chi_i$ is a fractional difference (L is the lag operator) of χ with a fractional differentiating parameter δ.[5] $\tau^{q=1}$ is time invariant and $\tau_m^{q=2}$ has a dimension of $M = 2$ regimes, identified by different levels of the threshold variable J_i (e.g., [47]). $J_i = ti_i$ is the inverse of volume weighted duration, which is estimated as the product of diurnally adjusted durations and $K(volume_i) = e^{\left(-\frac{volume_i - \overline{volume}}{\sigma_{volume}}\right)}$, where $\overline{volume}$ and σ_{volume} are the sample mean and standard deviation of volume, estimated per stock. J_i is an increasing variable of trading intensity. This specification models the inter-trade or volume weighted inter-trade waiting times as an infinite mixture of two q-Weibull distributions that are identified by a sole time-varying shape parameter, $\tau_i^{q=2}$.[6]

[5] This specification is preferred over more conventional specifications, such as in [29] due to the potentially long memory of durations. The rationale for selecting a long memory specification is primarily due to the objective pursued in this study. Eqq. (12)-(13) extract the information about the presence of different agent-types from the impact of their material actions. The impulse response functions, embedded in FIACD, are versatile enough to account for these actions. The parameter δ is the degree of decay that captures, in a sense, the life span of trading information. This allows dinstant past events to affect the conditional expectation of χ, $\psi_i = \mathbb{E}(\chi_i|\mathcal{F}_{i-1})$, and its conditional distribution, $f\left(\chi_i|\mathcal{F}_{i-1}; A_i\right)$. Variations in $f\left(\chi_i|\mathcal{F}_{i-1}; A_i\right)$, captured by $\boldsymbol{\tau}_m^q$, depend on ψ_i and, thus, on the long memory, i.e., δ, and its impact, i.e., ϕ, on ψ_i. This way, the identification of the different agents is done in a way that accounts for market reflexivity and past interactions of all agent-types. Additionally, the FIACD specification appears more computationally stable.

[6] When $\tau^{q=2} = 1$ the hazard function is flat matching the time-invariant arrival rate of uninformed agents, while when $\tau^{q=2} < 1$ ($\tau^{q=2} > 1$) the hazard function is decreasing (increasing) matching the characteristics of informed (technical, e.g., [46]) traders.

To articulate this more firmly, trading intensity, J_i, is employed as a classification variable, different levels of which are associated with differently shaped hazard functions and thus, with different agent-types. In line with previous literature (e.g., [25]) higher levels of trading intensity, i.e., $J_i^{q=2} > j_{m=2}^{q=2}$, are expected to be associated with decreasing hazard functions, i.e., $\tau_{m=2}^{q=2} < 1$, and therefore with a higher probability of this trade to have been initiated by an informed agent, captured by a higher value of $W_{m=2,i}^{q=2}$. Accordingly, $W_{m=2,i}^{q=2}$ is an estimate of the probability of the existence of private information and $W_{m=2,i}^{q=2}\lambda_0^{k=inf}(t,\tau_{m=2}^{q=2})$ is the conditional instantaneous probability of the arrival of an informed trader. Then, assuming that $\tau^{q=1} \to 1$ in Eq.(14), $\text{iVPIN}_i = \text{iVPIN}_{i:t_i \to t_i+\Delta t}$ can be written as:

$$\text{iVPIN}_{i:t_i \to t_i+\Delta t} = \frac{W_{m:\text{inf},i}^{q=2}(A_i((t_i+\Delta t)-t_i))^{\tau_{m:\text{inf}}^{q=2}}}{\sum_{m=1}^{3} W_{m,i}^{q=2}(A_i((t_i+\Delta t)-t_i))^{\tau_m^{q=2}}}$$

4.2 Empirical Estimation and Performance

In order to get comparable estimates, for both intensity-based and interval measures, the focus of the analysis is time, rather than volume. The objective is to create contemporaneous estimates, comparable across both frameworks. This is not an issue for the intensity-based iVPIN, but for the estimation of VPIN the time and length of interval are parameters of choice. Although fixed-volume buckets might perform better (e.g., [59]), the empirical setup here employs fixed-time buckets in order to create contemporaneous estimates. In particular, the sample, e.g., simulated (Section 5.3) and real data, is split into fixed-time buckets of different time lengths, $\text{bucketsize} = (1'', 5'', 15'', 30'', 1', 5', 15', 30', 60')'$, and all measures are estimated at the end of each bucket, at a time noted as t_{bucket}. This approach is an indirect way to provide robust estimates with respect to the length of the interval and the forecasting horizon.[7] Then VPIN and iVPIN are estimated using the same sampling frequency, over 10

[7] Previous literature (e.g., [59]) suggests that volume buckets might be more relevant in identifying private information because they identify the same magnitude of information, captured by a unit of volume. Then information is identified by the speed of volume accumulation. This approach would create timed buckets of variant time intervals, which would undermine comparability with iRP. Instead, the comparison here is based on fixed intervals for comparability reasons. The fixed-time intervals vary from 1" to 1h. The

lags, following the principle that: $\mathbb{E}\left(\frac{\alpha_{t_{bucket}}\mu_{t_{bucket}}}{(\alpha_{t_{bucket}}\mu_{t_{bucket}}+e_{t_{bucket}})}\right) \rightarrow \frac{\alpha\mu}{(\alpha\mu+e)}$. The difference lies in how each element is estimated. In particular:

VPIN. The benchmark metric for comparison is the conventional VPIN, which is estimated based on the principle ([20]) that $\mathbb{E}\left(\left\|V^B - V^S\right\|\right) \rightarrow \alpha\mu$ and $\mathbb{E}\left(\left\|V^B + V^S\right\|\right) \rightarrow \alpha\mu + e$. The aggregated buy $V^B \rightarrow V^B_{t_{bucket}}$ and sell $V^S \rightarrow V^S_{t_{bucket}}$ volumes are computed over different time intervals, i.e., bucketsize, with t_{bucket} marking the time at the end of each bucket, acting as a time identification. Then, $\mathrm{VPIN}_{t_{bucket}}$ is computed over a rolling window of $n = 10$ lags as: $\mathrm{VPIN}_{t_{bucket}} = \frac{\sum_{t_{bucket}-n}^{t_{bucket}}\left|V^B_{t_{bucket}} - V^S_{t_{bucket}}\right|}{\sum_{t_{bucket}-n}^{t_{bucket}} V_{t_{bucket}}}$.[8] In this metric, the most important element is the identification of trade initiation; a variable that is absent in raw data. This, has been shown to be an important issue (e.g., [59]), affecting the performance of VPIN, especially in combination with different sampling frequencies. The trade classification rules that exhibit superior performance are the "EMO" ([28]) trading classification rule and the Bulk Volume (hereafter BV: [24]) classification. The EMO classifies trades according to whether they are taken from the ask (buy) or bid (sell) price, while all other trades are classified according to the "tick" rule (e.g., [36]). In a misuse of the term, the estimate of VPIN using the EMO classification is called PIN, due to its resemblance to the lower frequency estimate of PIN.[9] In parallel, the BV classification de-

estimation of VPIN requires a rolling window of specific lag-structure and various studies suggest that the optimal length is data-driven. In the analysis below the VPIN is estimated using $n = 10$ lags and according to the bucketsize this might cover a period from 10" (still too long for algorithmic trading, but short enough for noise in trade direction to have a substantial impact in trade initiation identification) to 10h (longer than a day, which is long enough to reduce the impact of noise in trade initiation identification).

[8] The length of the rolling window of 10 is selected in order to facilitate the investigation of the performance of intensity-based measures in HFT. when bucketsize $= 1''$ the rolling window of 10 corresponds to 10". This interval is sufficiently short for algorithmic trading standards in order to evaluate whether the VPIN metric is of relevance at this sampling frequency, while it is also sufficiently long enough to avoid missing observations due to lack of data. On the opposite side, a rolling window of bucketsize $= 60'$ corresponds to a full trading day. Previous literature (e.g., [59]) shows that this interval provides reasonable estimates for VPIN, while it is still relevant to algorithmic trading, whose trading horizons rarely exceed one trading day.

[9] The original PIN employs a per trade classification rule and [22, 23] claim that one month worth of data is sufficient to estimate PIN. However, this would be rather unrealistic in HFT. Instead, VPIN is a more viable alternative, due to its HFT relevance (e.g., [27]). Following the idea of the "volume clock", unlike in the original PIN, the high-frequency metric employed here, estimates the order imbalance based on the total volume (rather than in the number of trades). This approach still considers the direction of trading derived from trade classification algorithms (like in the original PIN), but combines it with trading volume (like in VPIN). This provides a time variant estimate of PIN, which is preferred to other alternatives, such as the duration weighted trade imbalances of [20], due to its flexible sampling frequency.

rives trade initiation solely from price changes. In particular, using the notation in this paper, $V^B_{t_{bucket}}$ and $V^S_{t_{bucket}}$ are defined as $V^B_{t_{bucket}} = V_{t_{bucket}} Z\left(\frac{Price_{t_{bucket}} - Price_{t_{bucket}-1}}{\sigma_{\Delta Price}}\right)$ and $V^B_{t_{bucket}} = V_{t_{bucket}}\left(1 - Z\left(\frac{Price_{t_{bucket}} - Price_{t_{bucket}-1}}{\sigma_{\Delta Price}}\right)\right)$.

<u>iVPIN</u>. As an example of iRP, in Eq.(12), the conventional VPIN is compared to an intensity-based alternative, coined iVPIN. In order to create a comparable estimate, iVPIN is estimated on a per-trade basis, considering the inter-trade durations, i.e., $\chi_{i+1} = (t_{i+1} - t_i)$ for trades or $(t^*_{i+1} - t^*_i)$ for volume weighted durations, as the waiting time, i.e.: $\text{iVPIN}_{i:\chi_{i+1}} = \frac{W^{q=2}_{m:\text{inf},i}(A_i\chi_{i+1})^{\tau^{q=2}_{m:\text{inf}}}}{\sum_{m=1}^{3} W^{q=2}_{m,i}(A_i\chi_{i+1})^{\tau^{q=2}_m}}$. This provides a rather granular estimate of the expected level of VPIN, that can be, then, estimated at any desirable interval. VPIN is estimated for each bucketsize $= (1'' \, to \, 60')'$ and then an average over a time interval $n = 10$. The same approach is applied in the case of $\text{iVPIN}_{t_{bucket}}$, which is estimated as: $\text{iVPIN}_{t_{bucket}} = \frac{\sum_{t_{bucket}-n}^{t_{bucket}} \frac{\sum_i^{\#trades_{t_{bucket}}} \text{iVPIN}_i}{\#trades_{t_{bucket}}}}{n}$. For reference, when $\chi_{i+1} = (t_{i+1} - t_i)$, the intensity-based estimate is named iPIN, while, when $\chi_{i+1} = (t^*_{i+1} - t^*_i)$, it is named iVPIN.

<u>iVPIN vs VPIN</u>.After computing VPIN and iVPIN their performance is evaluated based on their forecasting ability on subsequent variance and the existence of UEE's. Following [3] this is evaluated based on the following regression:

$$Q_{t_{bucket}} = c_0 + c_1 Metric_{t_{bucket}-n} + \boldsymbol{cCV}_{t_{bucket}-n} + f.e. + \epsilon_{t_{bucket}} \tag{18}$$

where, $Q_{t_{bucket}} = (RV_{t_{bucket}}, UEE_{t_{bucket}})'$. The values are multiplied $x100$ in order to adjust the decimal places of the estimated coefficients. $Metric = (PIN, VPIN, iPIN, iVPIN)'$. $RV_{t_{bucket}}$ is the average realized volatility of each bucket over the time interval n. $UEE_{t_{bucket}}$ is the average number of UEE's: ([43]) of each bucket over the time interval n. [10] $\boldsymbol{CV} = (RV, spread, orders, averageduration)'$ is a collection of standard market microstructure variables that are introduced to control for varying market conditions and the sensitivity of VPIN to them (e.g., [14]). $f.e.$ is company fixed effects.

[10] UEE's are defined as periods that last less than 1,500ms, they follow a run (same trade direction) that exceeds 10 trades, during which prices change by more than 0.8% of the price at the beginning of the run.

4.3 Application: Real Data

The specification defined in Eq.(16) (as well as some extensions in Appendix B.2) is indicatively evaluated on a one-year, 2/1/2019-6/12/2019, sample of all constituents of Dow Jones Industrial Average (aka DOW30). The primary objective is to evaluate the performance of iRP against interval-based metrics, which would be better facilitated by a sample with minimal market-specific stylized factors or other types of trading biases. DOW30 consists of liquid, large-cap stocks and, from this perspective, is less likely to suffer from market specific biases. In addition, the constituents do not change during the sample period and, therefore, the trading activity is unlikely to be affected by portfolio rebalances. Also, the sample stops prior to the COVID-19 news and, thus, it should not be affected by it.

Concerning the data collection and manipulation process; information on all transactions is collected and for each transaction the associated date, time-stamp (millisecond), price ($) and trading volume (number of stocks) are recorded. The trade direction is not included in the source data and it is inferred by using the ”EMO” ([28]) trade classification rule. All observations outside the “normal trading hours” as well as the first transaction of each day (aggregated volume of the pre-opening session) have been omitted. According to [57] the nature of trading and the market participants have changed drastically with the technological advancements and this has direct implications on how information is diffused. The shift from fundamental to ”trading” information, which might render intensity-based metrics more adept, becomes increasingly relevant in the presence of algorithmic trading, which dominates the trading during the opening hours. Consequently, focusing on normal trading hours generates a less biased data sample. Furthermore, all trades with identical time stamp, price and trade initiation are considered as one segmented trade with aggregated volume. This accounts for passive splitting and mitigates the information loss of the trade classification algorithm. In addition, it reduces the proportion of trades with zero duration or zero price change, which, beyond the computational benefits, creates a sample that is focused on the time evolution of volume and price (or price change) withouth thinning the data. Moreover,

Table 2: Descriptive Statistics: Full Sample

	Full Sample #230,084,293				Simulation 1 asset 100 days x #100,000			
	Avg.	Min.	Max.	Std.	Avg.	Min.	Max.	Std.
Return	0	-9.76	9.77	0.04	0	-8.72	14.01	0.04
Volume	205.8	1	88,000	538.94	355.48	1	112,000	355.78
Duration	0.90	2.1E-5	4,500	2.34	0.85	0	3,856	1.99

Table 2 presents the descriptive statistics, i.e., the average (mean), the maximum (max), the minimum (min) and the standard deviation (std) of duration (in seconds), trading volume (in number of stocks) and price change (in $'s). The left panel presents the cross-sectional estimates of the statistics for the full sample that consist of all constituents of DJ30 (full table in Appendix). The left panel presents the cross-sectional estimates of the simulation that generates a calendar quarter worth of data (100 days, assuming 100,000 observations per day) the # sign is the count of observations per stock.

duration has been computed as the time between two consecutive trades, with one second being added to all observations for computational reasons, excluding the overnight period and has been diurnally adjusted (Engle and Russell, 1998). This results in a panel dataset of all filtered transactions of 30 firms with 230,084,293 unique observations.

The basic statistics of the final sample are presented in Table 2 (a more extensive, per-stock, presentation in Appendix A . The sample consists of rather liquid stocks like AAPL, to relatively less liquid stocks, like AXP, with an average duration of less than 1 second; around 0.9 seconds. The average volume per trade is just over 200 stocks; 205.8 and it is over-dispersed, standard deviation is 538.94, indicating a wide range of values. In addition, price changes exhibit some moderate variation at around 0, with a standard deviation of 0.04. These values are consistent with relevant literature, which shows that our sample is relatively homogeneous, but with adequate variation, in order to provide a sample with minimal trading biases or extreme events.

Table 3 presents the estimates of the parameters of the empirical specification in Table 1. The left panel reports the parameters of the conditional mean specification, i.e., Table 1, assumed to follow a fractionally integrated data generation process; $\psi_i = \omega + \beta\psi_{i-1} + (\chi_i - \beta\chi_{i-1}) - (\tilde{\chi}_i - \phi\tilde{\chi}_{i-1})$. $\tilde{\chi}_i = (1-L)^{\delta}\chi_i$ is a fractional difference (L is the lag operator)

Table 3: Estimation Results

	iPIN	*iVPIN*		*iPIN*		*iVPIN*	
				low	high	low	high
ω	0.6298	0.5305	q	1.3069		1.0144	
	(0.02)	(0.01)		(0.01)		(0.02)	
β	0.3073	0.5376	$(\tau\|\text{ti})$	1.1480	0.6376	1.1853	0.5945
	(0.03)	(0.02)		(0.02)	(0.02)	(0.02)	(0.02)
ϕ	0.4898	0.3695	*g(ti)*	1.0077		0.9951	
	(0.01)	(0.02)		(0.00)		(0.02)	
δ	0.1565	0.2475	*j(ti)*	1.1388		1.0639	
	(0.02)	(0.02)		(0.03)		(0.03)	

The left panel of Table 3 presents the estimation results for the conditional mean specification parameters, assuming a FI-ACD specification $\omega + \beta\psi_{i-1} + (\chi_i - \beta\chi_{i-1}) - (\tilde{\chi}_i - \phi\tilde{\chi}_{i-1})$. The right panel presents the distribution parameter estimates, assuming a $q-$Weibull distribution for χ, i.e., $f\left(\chi_i|\mathcal{F}_{i-1}\right) = (2-\tau^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right)$, where $A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}}/\psi_i\right]$, $\tau_i^{q=2} = \left(G_{m=1,i}^{q=2} - G_{m=2,i}^{q=2}\right)\tau_{m=1}^{q=2} + G_{m=2,i}^{q=2}\tau_{m=2}^{q=2}$ and $G_{m=2,i}^{q=2} = \left(1+e^{-g_{m=2}^{q=2}(J_i - j_{m=2}^{q=2})}\right)^{-1}$. The estimates are for the iPIN and iVPIN specifications in section 4.1 All estimates are cross-sectional averages, with standard deviations in (:).

of χ with a decaying parameter δ. δ is the long memory parameter and is estimated in advance using an ARFIMA$(0,\delta,0)$ specification. $\tilde{\chi}_i$ is the residuals of this specification. δ takes the value of 0.1565 (0.2475 for VPIN), suggesting a long memory ([41]), but not a fully integrated process. The values for β and ϕ indicate strong persistence, but at the same time satisfy the positivity constraints $\beta - \delta \leq \phi \leq \frac{2-\delta}{3}$ and $\delta\left(\phi - \frac{1-\delta}{2}\right) \leq \beta(\delta - \beta + \phi)$ of [13].

The right panel reports the estimates of the parameters of the distribution in Table 1, i.e., $f\left(\chi_i|\mathcal{F}_{i-1}\right) = (2-\tau^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right)$, where $A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}}/\psi_i\right]$, $\tau_i^{q=2} = \left(G_{m=1,i}^{q=2} - G_{m=2,i}^{q=2}\right)\tau_{m=1}^{q=2} + G_{m=2,i}^{q=2}\tau_{m=2}^{q=2}$ and $G_{m=2,i}^{q=2} = \left(1+e^{-g_{m=2}^{q=2}(J_i - j_{m=2}^{q=2})}\right)^{-1}$. The entropy parameter q is converging to 1 and this shows a convergence to a Weibull distribution. The smoothness parameters, $g_{m=2}^{q=2}$ noted as $g(ti)$'s, are very close to one, while the threshold values, $j_{m=2}^{q=2}$ noted as $s(ti)$'s, are very close to the unconditional means. This shows, a rather smooth transition from one regime to the other. These, data identified, regimes exhibit two distinct trading groups, as they are captured by the shape parameters, $\tau^{q=2}_{m=1}$ and $\tau^{q=2}_{m=2}$, noted as $\tau_{ti<j(ti)}$ (low) and $\tau_{ti>j(ti)}$ (high), respectively. $\tau_{ti<j(ti)}$'s are consistently higher than 1 and $\tau_{ti>j(ti)}$'s are consistently less than 1. This is consistent with previous literature (e.g., [46, 47, 48]) and indicates that higher trading intensity (i.e., $ti > j(ti)$) is associated with an increasing probability of informed trading. According to the estimates, when the threshold variable, J_i, takes values that are higher than 1.1388 (1.0639), the iPIN (iVPIN) metric indicates that there is an increasing probability that this trade was instigated by an informed trader. This is inferred by its sample-wide post trade impact that is reflected on a decreasing hazard function. The higher J_i is, the closer the shape parameter is to $\tau^{q=2}_{m=2}$ (0.6376 for PIN and 0.5945 for VPIN) and, thus, the sharper is the decreasing shape of the hazard function (indicating an accelerating market and, thus, a trade with high(er) post-trade impact), which is interpreted as more informative.[11] Heuristically, this indicates that higher trading intensity is associated with higher presence of private information. This finding is consistent with seminal studies in the market microstructure literature, e.g., [25], but the novelty in the metric proposed here is that it provides an empirical framework to assess the exact probability. In addition, the smoothness parameter is an indirect measure of how easy is to infer information from trading signals and, thus, it provides a direct estimate of market opacity.[12]

Furthermore, in order to provide a direct comparison between the intensity-based versus the interval-based metrics of PIN, their relative performance is tested in forecasting realized volatility ([59]) and UEE's ([43]). Table 4 presents the estimations results of Eq. (18) for variance and Table 5 for UEE's. In each table, the top panel presents the estimates of the parameter c_1, with t-statistics in () and it is followed by the R^2 and (M)ean (S)quared (E)rror

[11] [57] claims that algorithms use trading information as a noisy proxy for fundamental information and that speed is the new token of information. By extension, each trade is potentially informative and its ”information load” (e.g., [43]) can be assessed by its post-trade impact. This is the basic principle that associates a decreasing hazard function (i.e., accelerating post-trade impact) with information.

[12] A sharper (smoother) transition, captured by a higher (lower) value of the smoothness parameter, $g^{q=2}_{m=2}$, would indicate a clearer (less clear) distinction between the two regimes. Considering that the presence of different agent-types is latent information, a clearer (less clear) distinction would indicate greater (lower) market opacity because it is easier (more difficult) to extract latent information from observable signals.

Table 4: Real Data: Performance of metrics-Variance

	1”	5”	15”	30”	1’	5’	15’	30’	60’
					Estimates				
PIN	-0.0054	-0.0084	-0.0103	-0.0073	-0.0059	-0.0058	-0.0025	-0.0026	-0.0024
(t)	(1.44)	(74.27)	(64.22)	(56.63)	(-37.03)	(-34.24)	(-20.36)	(-3.63)	(-2.29)
R2	0.6688	0.5749	0.5249	0.4951	0.3238	0.3220	0.2726	0.2581	0.2436
MSE	0.0198	0.0163	0.0154	0.0132	0.0128	0.0120	0.0088	0.0068	0.0065
iPIN	0.6816	0.4112	0.3926	0.3239	0.2722	0.2326	0.1950	0.1306	0.0933
(t)	(124.44)	(76.17)	(66.62)	(61.91)	(40.64)	(37.30)	(22.01)	(10.60)	(8.51)
R2	0.6817	0.5819	0.5285	0.4980	0.3256	0.3229	0.2731	0.2646	0.2601
MSE	0.0160	0.0133	0.0130	0.0113	0.0111	0.0106	0.0080	0.0061	0.0052
VPIN	-0.0017	-0.0077	-0.0093	-0.0053	-0.0051	-0.0498	-0.0036	-0.0032	-0.0032
(t)	(-0.32)	(-52.26)	(-62.15)	(-48.83)	(-36.52)	(-32.14)	(-20.01)	(-8.46)	(-6.00)
R2	0.6683	0.5740	0.5231	0.4947	0.3233	0.3212	0.2721	0.2643	0.2587
MSE	0.0257	0.0187	0.0165	0.0147	0.0130	0.0121	0.0091	0.0068	0.0056
iVPIN	0.8395	0.7551	0.6949	0.4712	0.4216	0.3743	0.2758	0.2213	0.2105
(t)	(133.90)	(92.07)	(55.75)	(52.53)	(50.47)	(40.58)	(26.93)	(12.81)	(9.21)
R2	0.6848	0.5849	0.5312	0.5007	0.3281	0.3252	0.2753	0.2668	0.2609
MSE	0.0158	0.0124	0.0123	0.0110	0.0102	0.0100	0.0074	0.0061	0.0051
					MSE				
PIN/iPIN	1.2377	1.2277	1.1863	1.1660	1.1466	1.1308	1.1017	1.1135	1.2512
VPIN/iVPIN	1.6330	1.5074	1.3401	1.3370	1.2710	1.2071	1.2295	1.1109	1.0951
iPIN/iVPIN	1.0138	1.0725	1.0549	1.0314	1.0907	1.0526	1.0782	1.0019	1.0239
PIN/VPIN	0.7684	0.8735	0.9339	0.8995	0.9839	0.9861	0.9662	1.0041	1.1699
PIN/iVPIN	1.2548	1.3167	1.2515	1.2026	1.2506	1.1903	1.1879	1.1155	1.2811
					R2				
PIN/iPIN	0.9810	0.9879	0.9933	0.9940	0.9945	0.9972	0.9981	0.9755	0.9368
VPIN/iVPIN	0.9760	0.9813	0.9847	0.9878	0.9854	0.9878	0.9883	0.9907	0.9918
iPIN/iVPIN	0.9955	0.9950	0.9947	0.9946	0.9924	0.9928	0.9919	0.9916	0.9969
PIN/VPIN	1.0007	1.0016	1.0035	1.0009	1.0016	1.0022	1.0017	0.9764	0.9417
PIN/iVPIN	0.9955	0.9950	0.9947	0.9946	0.9924	0.9928	0.9919	0.9916	0.9969

Table 4 presents the estimation results for Eq. (18), where the dependent variable is Realized volatility. All estimations inclued the same control variables and company fixed effects. The top panel presents the estimates of the coefficients with t-stats in (), as well as the adjusted R^2 and the (M)ean (S)quared (E)rror (MSE). The bottom two panels report the ratios of MSE and R^2 for the pairs indicated on the left.

(MSE). All estimations include asset fixed effects and a set of microstrure controlled variables, the estimates of which are not reported for brevity. Each section named, PIN, iPIN, VPIN and iVPIN are separate estimations with each metric being considered independently of the others. The comparison is based on R^2 and the MSE and the bottom panel presents the ratios for direct commparison. Finally, both tables are organized into columns according to the interval frequency, bucketsize $= (1'', 5'', 15'', 30'', 1', 5', 15', 30', 60')'$.

Table 5: Real Data: Performance of metrics-UEE's

	1”	5”	15”	30”	1’	5’	15’	30’	60’
Estimates									
PIN	0.0352	0.0542	0.0666	0.0473	0.0381	0.0372	0.0159	0.0166	0.0158
(t)	(4.84)	(36.27)	(31.36)	(27.65)	(18.08)	(16.72)	(9.94)	(7.77)	(4.90)
R2	0.0679	0.0576	0.0524	0.0496	0.0324	0.0324	0.0273	0.0258	0.0243
MSE	0.5025	0.4976	0.0492	0.4312	0.0321	0.0319	0.2834	0.0255	0.0268
iPIN	0.0573	0.0346	0.0330	0.0272	0.0229	0.0196	0.0164	0.0110	0.0079
(t)	(49.45)	(39.23)	(34.31)	(31.89)	(20.93)	(19.21)	(14.34)	(9.91)	(7.96)
R2	0.0686	0.0586	0.0532	0.0501	0.0328	0.0325	0.0275	0.0266	0.0262
MSE	0.4944	0.4618	0.4516	0.3943	0.3865	0.3669	0.2764	0.2120	0.1792
VPIN	0.0367	0.1613	0.1955	0.1111	0.1075	1.0465	0.0761	0.0671	0.0664
(t)	(3.39)	(35.55)	(42.28)	(33.22)	(24.84)	(21.87)	(13.61)	(8.76)	(6.21)
R2	0.0648	0.0573	0.0522	0.0493	0.0321	0.0322	0.0271	0.0263	0.0257
MSE	0.5763	0.0583	0.0527	0.0501	0.0326	0.0326	0.0276	0.0254	0.0229
iVPIN	0.0598	0.0433	0.0398	0.0270	0.0242	0.0215	0.0158	0.0127	0.0121
(t)	(51.07)	(41.29)	(32.20)	(27.17)	(19.07)	(16.55)	(15.17)	(13.84)	(10.57)
R2	0.0696	0.0594	0.0540	0.0509	0.0333	0.0330	0.0280	0.0271	0.0265
MSE	0.4726	0.4314	0.4291	0.3828	0.3552	0.3498	0.2577	0.2125	0.1767
MSE									
PIN/iPIN	1.0164	1.0766	0.1104	1.0955	0.0833	0.0871	1.0301	0.1208	0.1496
VPIN/iVPIN	1.2196	0.1347	0.1241	0.1303	0.0926	0.0927	0.1058	0.1202	0.1314
iPIN/iVPIN	1.0462	1.0714	1.0459	1.0281	1.0869	1.0479	1.0676	0.9954	1.0195
PIN/VPIN	0.8719	8.5642	0.9303	8.6444	0.9787	0.9843	10.3904	1.0003	1.1606
PIN/iVPIN	1.0634	1.1535	0.1155	1.1262	0.0906	0.0913	1.0997	0.1202	0.1526
R2									
PIN/iPIN	0.9893	0.9813	0.9843	0.9843	0.9898	0.9897	0.9979	0.9660	0.9276
VPIN/iVPIN	0.9311	0.9591	0.9657	0.9629	0.9669	0.9683	0.9718	0.9693	0.9695
iPIN/iVPIN	0.9861	0.9856	0.9853	0.9852	0.9831	0.9834	0.9825	0.9822	0.9875
PIN/VPIN	1.0478	1.0084	1.0043	1.0071	1.0063	1.0051	1.0088	0.9789	0.9447
PIN/iVPIN	0.9861	0.9856	0.9853	0.9852	0.9831	0.9834	0.9825	0.9822	0.9875

Table 5 real presents the estimation results for Eq. (18), where the dependent variable is UEE. All estimations inclued the same control variables and company fixed effects. The top panel presents the estimates of the coefficients with t-stats in (), as well as the adjusted R^2 and the (M)ean (S)quared (E)rror (MSE). The bottom two panels report the ratios of MSE and R^2 for the pairs indicated on the left.

The empirical findings using both metrics are rather consistent and provide evidence that the intensity-based metrics perform notably better, especially in higher sampling frequencies. The first observation is that iVPIN and iPIN exhibit a higher R^2 and a lower MSE than VPIN and PIN, in this order, in all sampling frequencies. This is more pronounced in higher sampling frequencies, espcially when the interval is closer to 1” (more relevant for HFT). This highlights that the initial motivation of this paper that interval-based measures might not be adequately adopted for HFT, even when the volume-clock is considered in VPIN. The ratios reported at the bottom of the two tables show that in lower frequencies, e.g., 30' or 60', the performance differences persist, but all the metrics seem to converge. In contrast, the biggest differences are observed in higher sampling frequencies, e.g., 1" to 15", highlighting that, in HFT, the aggregation process itself, i.e., arrival rates and hazard functions, constitute a stronger signal, compared to the aggregated values, e.g., aggregated trade direction or price change. This finding is consistent cross-sectionally and provides a first evidence in favour of the underlying concept in this study, that in HFT, where algorithms act in sub-human attention speeds, interval estimates are inapt in capturing the properties of the data. Even worse, these findings are a first evidence that the impact of the noisy signals is higher in higher sampling frequencies and, thus, these metrics exhibit higher MSE's. Instead, the modeling of the arrival rates with conditional intensities (hazard functions) performs notably and consistently better, because it focuses on the granular properties of the data. Conditional intensities are a mathematical tool to model instantaneous probabilities, and for this reason, they describe better the aggregation properties of the data.

In addition, trading volume enhances the performance of all metrics. In intensity-based metrics the volume enhanced iVPIN outperforms consistently the iPIN that is based solely on the arrival rate of trades, rather than of volume. In interval-based metrics, VPIN exhibits some significant noise in high sampling frequencies, 1" to 1', but it outperforms PIN in intervals that exceed 15'. This highlights that the initial intuition behind VPIN that the noise present in the trade direction might be mitigated by focusing on the volume-clock, is toward the right direction. However, classification according to trade direction is still noisy, independently on whether is based on trade direction (PIN) or price change (VPIN). Instead, focusing on the aggregation properties of trading volume is an approach that exhibits a consistently higher performance, that persist even at lower trading frequencies.

5 Theoretical Properties of the Intensity-Based Estimator

Section 4 provides some empirical evidence in favour of the intensity-based estimates, especially when they are volume enhanced. However, in order to verify that they are not circumstantial due to sampling bias, Section 5 investigates the statistical theory of the plug-in estimator for the iRP, through the limit theory of the maximum likelihood estimator for the parameters that appear in the specification. Section 5.1 provides (pseudo-) consistency considerations, showing the adequacy of the application (e.g., iPIN and iVPIN) of iRP in capturing the true PIN. Section 5.2 discusses rates, asymptotic distributions and subsequently issues of asymptotic inference and acts as a base for a comparison between the interval-based and the intensity-based estimators. Finally, Section 5.3 discusses the impact of mis-specification in the intensity-based metrics (iRP) through a Monte-Carlo simulation. The discussion in Section 5 evolves around private information, but it is easily expandable into other agent-types (e.g., Appendix C.2).

In more detail, the definition and the limit theory of the maximum likelihood estimator (MLE) for the parameters of interest are presented and derived in this section. The statistical model at hand is allowed to be misspecified. The point processes involved assume their values in an interval of the form $[0, T]$. The asymptotics operate as $T \to +\infty$. $\rightsquigarrow$ denotes convergence in distribution.

The latent conditional duration process $(\psi_i)_{i\in\mathbb{Z}}$ is assumed to be a solution of a stochastic recurrence equation (SRE) of the form $\psi_i = \Psi(\theta_0, \psi_{i-l}, \chi_{i-j}, l = 1, \ldots, q, j = 1, \ldots, p)$, where Ψ is a real function, θ_0 is an unknown value of a Euclidean parameter θ that belongs to some known $\Theta \subseteq \mathbb{R}^l$, while p is allowed to assume the value $+\infty$ to incorporate ARCH(∞)-type of elements associated for example with the FIACD model used in the empirical section. We also have that for each $i \in \mathbb{Z}$, χ_{i-l}, $\varepsilon_{i-l}, \mathbf{J}_{i-l}$, $l \geq 1$, are measurable w.r.t. $\mathcal{F}_{i-1}$. The parameters of interest are collected in the Euclidean vector $\phi := (\theta^T, \tau^T)^T$, that assumes its values in a known to the researcher subset Φ of the Euclidean space $\mathbb{R}^{l+3MQ}$, where $\tau := \mathrm{vec}(\tau^q_m, g^q_m, j^q_m)^{q=1,\dots,Q}_{m=1,\dots,M}$. The researcher has at her disposal the sample realizations of the observable processes $y := (\chi_i, \mathbf{J}_i)_{i=1,\dots T}$, with $T \geq \max(p^\star, q)$, for $p^\star$ a finite truncation of any non trivial ARCH(∞) component, and selects a-potentially random-initialization $\hat{\psi}_0(\theta)$, $\theta \in \Theta$, constructing a recurrent a filter of the latent conditional duration defined by:

$$\hat{\psi}_i(\theta) := \begin{cases} \hat{\psi}_0(\theta),\ i = 0, \\ \Psi^\star((\theta; \hat{\psi}_{i-l}, x_{i-j}, l = 1, \dots, \min(i, \max(p^\star, q)),\ j = 1, \dots, \min(i, p^\star)),\ 0 < i \leq T, \end{cases}$$

where for the modified SRE that appears in the previous display we have that $\Psi^\star((\theta; \hat{\psi}_{i-l}, \chi_{i-j}, l = 1, \dots, \max(p^\star, q),\ j = 1, \dots, \min(i, p^\star)) := \Psi((\theta; \hat{\psi}_{i-l}, {}^{\chi_{i-j}}/_{\hat{\psi}_{i-j}}, l = 1, \dots, \min(i, q),\ j = 1, \dots, \min(i, p^\star))$. Then, the log-likelihood function is defined by $\ell_T(y; \phi) := {}^{1}/_{T} \sum_{i=1}^{T} \ell_i(\hat{\psi}_i, \tau_i; \phi)$, where the likelihood contributions are $\ell_i(\hat{\psi}_i, \tau_i; \phi) := \ln f(\chi_i; \hat{\psi}_i, \tau_i; \phi))$. The MLE, say ϕ_T, is defined via the variational problem

$$\ell_T(y; \phi_T) \leq \inf_{\Phi} \ell_T(y; \phi) + \epsilon_T,$$

with ϵ_T almost surely non-negative that admits the role of optimization error. The specification appearing in Table 1 is readily conformable to the above.

5.1 (Pseudo-)Consistency of the new estimator

The following assumption framework ensures existence of the estimator by standard arguments. We skip the details and focus on the issue of (pseudo-)consistency. The framework moreover ensures the existence of an approximating likelihood function with stationary and ergodic contributions, so that locally uniform versions of the ergodic LLN are applicable (or more generally the former almost surely approximates the latter w.r.t. hypo-convergence). The assumptions also posit that the limiting likelihood is uniquely minimized at a parameter value $\phi^\star$ that implies the minimization of the KL divergence between the true conditional (on the algebra $\mathcal{F}_i)_i$ density of χ_i and $f(\chi_i;\psi_i,\tau_i;\phi^\star)$. This allows for the consideration of the limiting behavior of the MLE even in cases of model misspecification. To be precise we consider the following assumptions:

A1: *The joint process $(\chi_i, \mathbf{J}_i)_{i\in\mathbb{Z}}$ is stationary and ergodic. The density of χ_i conditionally on $\mathcal{F}_{i-1}$ exists and has an integrable logarithm.*

Remark 1. Stationarity and ergodicity for $(\chi_i)_{i\in\mathbb{Z}}$ would follow from stationarity and ergodicity of $(\varepsilon_i)_{i\in\mathbb{Z}}$ and conditions that ensure existence and uniqueness (up to modification) of a solution to the Ψ-SRE defined via an almost sure limit of backward substitutions (see for example Ch.2 of [65]). If the specification appearing in Table 1 is correct, this would follow as long as $\max(|\beta_0|,|\phi_0|) < 1$ and $\delta_0 \in [0, 1/2)$, where β_0, ϕ_0, and δ_0 denote the unique true values for the auto-regressive and the fractional differencing parameters respectively-see the assumption that follows (see for example [41]). If the remaining parts of the process satisfy SREs analogous considerations would suffice. If the elements of the remaining parts are fixed (w.r.t. i) measurable transformations of underlying stationary and ergodic processes, stationarity and ergodicity would also be the case.

A2: *There exists a $\phi^\star \in \Phi$ such that $\mathbb{E}(\ell_0(\phi^\star)) > \mathbb{E}(\ell_0(\phi))$ for all $\Phi \ni \phi \neq \phi^\star$.*

Remark 2. This is an identification condition for the (pseudo-) true value of the parameter involved. In the case of the specification in Table 1, and if the model is well-specified-this follows from the fact that the conditional distribution of χ_i has a density (see for example Par. 5.4.2 of [65]). In the case of misspecification, the previous would also suffice due to Proposition 2.3 of [42].

In what follows, $\Theta^\star$ denotes an arbitrary compact subset of Θ.

A3: *Suppose that: (i). $\mathbb{E}(\sup_{\theta\in\Theta^\star} |\Psi^\star(\cdot)|) < +\infty$. (ii). $\Psi^\star(\theta;\psi_{i-l}, x_{i-j}, l = 1,\ldots,\max(p,q), j = 1,\ldots,p)$ is almost surely Lipschitz continuous in $(\psi_{i-l},\ l = 1,\ldots,\max(p,q))$, with Lipschitz coefficient $\Lambda_i(\theta)$, and such that, (a). the map $\Theta^\star \ni \theta \to \Lambda_0(\theta)$ is almost surely continuous and, (b). $\mathbb{E}(\sup_{\theta\in\Theta^\star} \log^+ \Lambda_0(\theta)) < 0$, where $\log^+$ is the positive part of the logarithmic* *function.*

Remark 3. The assumption implies the continuous invertibility (see for example Par. 3 and Prop. 3.1 of [9]) of the filter $(\hat{\psi}_i(\theta))$; this is equivalent to the existence of a stationary and ergodic process (say $(f_i(\theta))_{i\in\mathbb{Z}}$), that approximates appropriately fast, almost surely, and (locally) uniformly over (δ, θ), the original filter as $i \to \infty$. In the specification of Table 1, it is ensured whenever $\max(|\beta_0|, |\phi_0|)$ is bounded below 1. For more complicated filters this may not be the case (see Par. 6 of [9]).

A4: $\Phi^\star$ *denotes any compact subset of* Φ *such that if* $\Phi^\star \ni \phi = (\theta, \eta)$*, then* $\theta \in \Theta^\star$*. Suppose that there exists a stationary non-negative process process* $(m_i)_{i\in\mathbb{Z}}$*, with* $\mathbb{E}(\log^+ m_0) < +\infty$*, such that almost surely* $\sup_{\phi\in\Phi} |\ell_i(\hat{\psi}_i, \tau_i; \phi) - \ell_i(f_i, \tau_i; \phi)| \leq m_i \sup_{\theta\in\Theta^\star} |\hat{\psi}_i - \psi_i|$ *for any* $i \in \mathbb{N}$*, where* $f_i(\theta)$ *as in Remark 7.*

Remark 4. Given the continuous invertibility, the assumption allows the approximation of the likelihood function by a stationary and ergodic version that is constructed via the limiting filtering process $(f_i)_{i\in\mathbb{Z}}$. In the specification of Table 1 it holds when the elements of the (τ_i) process are almost surely bounded from above and away from zero, x_0 has a logarithmic moment (see also 1), and f_i is uniformly over θ bounded away from zero (see [41]).

A5: *The elements of* τ_0 *are almost surely continuous and bounded on* Ψ*, and* $\mathbb{E}(\sup_{\Phi^\star} \max(\ell_0(x_0; f_0, \tau_0; \phi), 0)) < +\infty$*.*

Remark 5. Given the stationary and ergodic version of the likelihood function, the assumption implies the applicability of the locally uniform version of Birkhoff's LLN so that the function converges almost surely to its expectation which is well defined. Similarly to the previous remark, in the specification of Table 1, or more generally in models where the (q)-Weibull density is used, it holds whenever the elements of the (τ_i) process are almost surely bounded from above and away from zero, and x_0 has enough moments; moment orders that approximate from above the essential supremum of the τ_0 suffice.

A1-**A5** imply then pseudo consistency:

Theorem 1. *Suppose that **A1**-**A5** hold, there exists a gamma* > 1 *such that* $\gamma^T|p^\star - p| \to 0$*, and* $\epsilon_T \to 0$*,* $\mathbb{P}$ *a.s. Then, (i). the expectation of the Kullback-Liebler divergence between the density in **A1** and* $f(x_i; \psi_i, \tau_i; \phi)$ *is well defined, and uniquely minimized at* $\phi^\star$*. (ii).* $\phi_T \to \phi^\star$*,* $\mathbb{P}$ *a.s. for any* $\hat{\psi}_0$*.*

The existence of a $\gamma > 1$ that validates the condition $\gamma^T|p^\star - p| \to 0$ in the case of the specification of Table 1 is ensured by the meromorphic continuation of the Barnes' Zeta function (see [61]) and the fact that that δ is not allowed to lie outside $[0, 1/2)$.

The result along with the idenification Assumption **A2** imply that the pseudo-true value $\phi_\star$ has a variational characterization; it is the unique minimum of the expected Kullback-Liebler divergence (see for example [2]) between the conditional distributions that appear in the statistical moment at hand, and the DGP distribution of χ_i.

When the model $\{f(x_i; \psi_i, \tau_i; \phi); \phi \in \Phi\}$, is well-specified, the previous imply that there exists some $\phi_0 \in \Phi$ such that the density in **A1** is $f(x_i; \psi_i, \tau_i; \phi_0)$. Then necessarily $\phi^\star = \phi_0$:

Corollary 1. *If the model is well-specified then* $\phi_T \to \phi_0$*,* $\mathbb{P}$ *a.s. for any* $\hat{\psi}_0$*.*

We complete this section with the issue of the strong approximation of the iPIN by its estimator based on the MLE and the conditional duration filter. Specifically, given the the cumulative hazard functions of the characteristics employed in the analysis, the iPIN at time t, given $\mathcal{F}_s$, say $\text{iPIN}_{t,s}(\phi_T)$, can be extracted by evaluating eq.(16) at ϕ_T as well as at the filter $\hat{\psi}_{s+1}(\phi_T)$. The following result is easily established via the CMT and Corollary 1:

Proposition 1. *Suppose that (i). assumptions **A1**-**A5** hold and* $\epsilon_T \to 0, \mathbb{P}$ *a.s., (ii). each element of* τ_i *is almost surely continuous in* ϕ*, (iii). the cumulative hazard functions employed are continuous functions of the shape and scale parameters, (iv)* $\mathbb{E}(\log^+ \sup_\theta f_0(\theta)) < +\infty$ *and (v). the statistical model is well specified. Then as* $s + 1 \leq T \to \infty$*,* $|\text{iPIN}_{t,s}(\phi_T, \hat{\psi}_{s+1}) - \text{iPIN}_{t,s}(\phi_0)| \to 0$*,* $\mathbb{P}$ *a.s. conditionally on* $\mathcal{F}_s$*.*

The continuity properties for the conditional shape parameters as well as condition (iv) of the proposition, hold trivially for the specification in Table 1. The compactness of Φ implies a uniform integrability argument that in turn implies that $\mathrm{iPIN}_{t,s}(\phi_T, \hat{\psi}_{s+1})$ converges to $\mathrm{iPIN}_{t,s}(\phi_0)$ in the L^2 mode, conditionally on $\mathcal{F}_s$. An analogous result holds for iVPIN. The same is true for the smoothed versions of both the PIN and VPIN estimators that appear in the previous paragraph, for fixed n. All those results are useful in the following paragraph.

5.2 Rate of Convergence and Weak Gaussian Approximation

Given the (pseudo-)consistency results of Theorem 1 for the MLE, we complete the limit theory by deriving its rate of convergence and a subsequent Gaussian approximation in distribution. We proceed using the classical analysis; under adequate differentiability, the asymptotics of the first order conditions for the optimization of the likelihood are derived. Similarly to the case of consistency, issues of invertibility for the SREs formed via differentiation of the filters that appear in the likelihood emerge. We deal with suchlike issues, by completing our set of assumptions as follows:

B1: *$\phi^\star$ lies in the interior of Φ.*

Remark 6. This enables the w.h.p. use of f.o.c.s. for the analysis. It can be easily discarded (see for example [4]) if-among others-the local parameter space has for example the structure of a convex cone. Even though the assumption essentially precludes the weak dependence case ($\delta_0 = 0$) in the model appearing in Table 1-if well specified, we do not pursue this generalization to avoid clutter.

B2: *There exists an open neighborhood, say $B_{\phi^\star}$, of $hi^\star$ such that: (i) $\Psi^\star$ is twice continuously differentiable w.r.t. (θ, ψ) on $B_{\phi^\star} \times \mathbb{R}^p$, for almost every value of its remaining arguments. τ_0 is twice continuously differentiable w.r.t. ϕ on $B_{\phi^\star}$ for almost every value of its remaining arguments. (ii). $\Psi^\star_{\partial\theta}$ denotes the SRE obtained by recursive differentiation of $\Psi^\star$ w.r.t. θ, partially via the chain rule through the derivatives of its ψ arguments w.r.t. θ. Then $\mathbb{E}(\sup_{\phi \in B_{\phi^\star}} ||\Psi^\star_{\partial\theta}(\cdot)||) < +\infty$, where $|| \cdot ||$ denotes the Euclidean norm. (A). $\Psi^\star_{\partial\theta}$ is almost surely Lipschitz continuous in $(\partial_\theta \psi_{i-l},\ l = 1, \ldots, \max(p, q))$, with Lipschitz coefficient $\Lambda^{(\partial\theta)}_i(\theta)$, and such that, (B). the map $B_{\phi^\star} \ni \theta \to \Lambda^{(\partial\theta)}_0(\theta)$ is* *almost surely continuous and, (C).* $\mathbb{E}(\sup_{\theta \in B_{\phi^\star}} \log^+ \Lambda_0^{(\partial\theta)}(\theta)) < 0$. *(iii). Analogously,* $\Psi^\star_{\partial\theta\partial\theta^T}$ *denotes the SRE obtained by recursive differentiation of* $\Psi^\star_{\partial\theta}$ *w.r.t.* θ*, partially via the chain rule through the derivatives of its* ψ *and* $\partial_\theta \psi$ *arguments w.r.t.* θ*. Then* $\mathbb{E}(\sup_{\phi \in B_{\phi^\star}} ||\Psi^\star_{\partial\theta\partial\theta^T}(\cdot)||) < +\infty$*, where* $|| \cdot ||$ *denotes the Frobenius norm. (A).* $\Psi^\star_{\partial\theta\partial\theta^T}$ *is almost surely Lipschitz continuous in* $(\partial_\theta \partial\theta^T \psi_{i-l},\ l = 1, \ldots, \max(p, q))$*, with Lipschitz coefficient* $\Lambda_i^{(\partial\theta\partial\theta^T)}(\theta)$*, and such that, (B). the map* $B_{\phi^\star} \ni \theta \to \Lambda_0^{(\partial\theta\partial\theta^T)}(\theta)$ *is almost surely continuous and, (C).* $\mathbb{E}(\sup_{\theta \in B_{\phi^\star}} \log^+ \Lambda_0^{(\partial\theta\partial\theta^T)}(\theta)) < 0$.

Remark 7. The assumption implies among others the continuous invertibility of the filter $(\hat{(\psi)}_i(\theta))$ derivatives. In the specification of Table 1, it follows whenever $\max(|\beta_0|, |\phi_0|)$ is bounded below 1.

B4: *There exists a stationary non-negative process process* $(m_{\partial i})_{i \in \mathbb{Z}}$*, with* $\mathbb{E}(\log^+ m_{\partial 0}) < +\infty$*, such that almost surely* $\sup_{\phi \in B_{\Phi_\star}} ||\partial_\theta \ell_i(\hat{\psi}_i, \partial_\theta \hat{\psi}_i, \tau_i; \phi) - \partial_\theta \ell_i(\psi_{\theta i}, \partial_\theta \psi_{\theta i}, \tau_i; \phi)|| \leq m_i \sup_{\theta \in B(\phi^\star})(|\hat{\psi}_i - \psi_i| + ||\partial_\theta \hat{\psi}_i - \partial_\theta \psi_i||)$ *for any* $i \in \mathbb{N}$*, and,* $\sup_{\phi \in B_{\Phi_\star}} ||\partial_\theta \partial_\theta \partial^T_{.} \ell_i(\hat{\psi}_i, \partial_\theta \hat{\psi}_i, \partial_\theta \partial^T_\theta \hat{\psi}_i, \tau_i, \partial_\eta \tau_i; \phi) - \partial_\theta \partial_{.} \ell_i(\psi_{\theta i}, \partial_\theta \psi_{\theta i}, \partial_\theta \partial^T_\theta \hat{\psi}_i, \tau_i, \partial_\eta \tau_i; \phi)|| \leq m_i \sup_{\theta \in B(\phi^\star})(|\hat{\psi}_i - \psi_i| + ||\partial_\theta \hat{\psi}_i - \partial_\theta \psi_i|| + ||\partial_\theta \partial_{\theta^T} \hat{\psi}_i - \partial_\theta \partial_{\theta^T} \psi_i||)$ *for any* $i \in \mathbb{N}$.

Remark 8. Again, given the continuous invertibility, the assumption allows the approximation of the score and the Hessian of the likelihood function by a stationary and ergodic version that is constructed via the limiting filtering process and its derivatives. For the specification appearing in Table 1, it holds whenever the elements of the (τ_i) process are almost surely bounded from above and away from zero, x_0 has a logarithmic moment (see also 1), and f_i is uniformly over θ bounded away from zero-the latter holds trivially in the particular example.

B5: $\mathbb{E}(\sup_{\phi \in B_{\phi^\star}} \|\partial_\phi \ell_0)\| + \mathbb{E}(\sup_{\phi \in B_{\phi^\star}} \|\partial_\phi \partial_{\phi^T} \ell_0)\|) < +\infty$*, and for some* $\delta > 0$*,* $\mathbb{E}(\|\partial_\phi \ell_0(\phi^\star)\|^{1+\delta}) < +\infty$*, for the stationary and ergodic versions of the derivatives. Furthermore, the elements of the stationary and ergodic version of the Hessian are linearly algebraically independent.*

Remark 9. The assumption implies the applicability of the locally uniform version of Birkhoff's LLN on the stationary and ergodic version of the Hessian, the identification of the limiting focs via dominated convergence, and-in conjunction with **B1** and **B3** the applicability of the aforementioned CLT. Similarly to the previous remark, in the specification of Table 1, it holds whenever the elements of the (τ_i) process are almost surely bounded from above and away from zero, and x_0 has enough moments; orders that approximate from above the essential supremum of the τ_0 on the power of $1+\delta$ suffice.

Utilizing the totality of our assumption framework along with some control of the rate at which the optimization error converges to zero, we obtain the following result-there $\rightsquigarrow$ denotes convergence in distribution:

Theorem 2. *Under the premises of Theorem 1, and if moreover* ***B1-B5*** *hold and* $\sqrt{T}\epsilon_T \rightsquigarrow 0$, *then*

$$\sqrt{T}(\phi_T - \phi^\star) \rightsquigarrow N(\mathbf{0}, (\partial_\phi \partial_{\phi^T} \ell_0(\phi^\star))^{-1}(\partial_\phi \ell_0(\phi^\star)\partial_\phi \ell_0^T(\phi^\star))(\partial_\phi \partial_{\phi^T} \ell_0(\phi^\star))^{-1}),$$

for the stationary and ergodic versions of the associated derivatives.

Again the existence of a γ that ensures the exponentially fast approximation of the part of the truncated filter that depends on the derivatives w.r.t. δ of the ARCH(∞) component is ensured by the meromorphic continuation of the Barnes' Zeta function (see [61]), the fact that δ is not allowed to lie outside a compact subset of $(0, 1/2)$, and the asymptotic representation of the series' coefficients as $Cj^{1-\delta}$ for some $C > 0$ independent of j, δ.

The limit theory involves a standard rate and asymptotic normality with the usual sandwich form for the asymptotic variance. Consistent estimators of the terms that appear in there can be easily obtained via the non-stationary versions of the derivatives evaluated at the MLE, due to consistency and Assumptions **B1**, **B3-B5**. This can be useful for the construction of Wald-type tests for parameters of interest. When the statistical model is well-specified then due to **B1-B5** and dominated convergence, the information matrix equality yields-as expected:

Corollary 2. *Under the premises of Theorem 2 and if the statistical model is well specified, then:*

$$\sqrt{T}(\phi_T - \phi_0) \rightsquigarrow N(\mathbf{0}, (\partial_\phi \ell_0(\phi_0) \partial_\phi \ell_0^T(\phi_0))^{-1}).$$

We close this section with an application of the above and the Delta method to the derivation of the limit theory of iPIN. In what follows d_W denotes any metric that metrizes weak convergence-see for example Par. 1.12 of [68]:

Proposition 2. *Under the premises of Theorem 2 and if the cumulative hazard functions employed are continuous functions of the shape and scale parameters, and the statistical model is well specified, then as* $s+1 \leq T \to \infty$*,* $d_W(|\sqrt{T}(\mathrm{iPIN}_{t,s}(\phi_T, \hat{\psi}_{s+1}(\phi_T)) - \mathrm{iPIN}_{t,s}(\phi_0)), N(0, \partial_\phi \mathrm{iPIN}(\phi_0)^T (\partial_\phi \ell_0(\phi_0) \partial_\phi \ell_0^T(\phi_0))^{-1} \partial_\phi \mathrm{iPIN}_{t,s}(\phi_0))| \to 0$*,* $\mathbb{P}$ *a.s. conditionally on* $\mathcal{F}_s$*.*

The latter can be useful for the construction of confidence sets. An analogous derivation obviously holds for the estimated iVPIN.

The construction of confidence sets for the plug-in estimators of iPIN and iVPIN can be also performed via Monte Carlo methods, due to the parametric nature of the statistical model at hand-when this is well specified. Specifically, the limiting (unconditional) variance of the estimators, can be consistently estimated via the MC empirical variance of the resulting iPIN (or iVPIN) estimator, when the DGP is evaluated at the originally estimated parameters. This is due to the CMT, the locally (w.r.t. ϕ) uniform convergencies mentioned above, and the consistency of MLE. A similar (yet possibly less accurate) approximation can be available from the cross sectional averaging of the iPIN (or iVPIN) estimators, when the cross sectional DGPs are similar, and satisfy some form of exchangeability property, or more generally invariance of the underlying joint distributions under groups of transformations (see for example [6]).

5.3 Misspecification and Trade Arrival Data Contamination: Some Monte Carlo Evidence

The predictive regressions appearing in Section 4.3, are easily interpretable when both the regression model and the specification appearing in Table 1 are well-specified. Then the plug-in estimators for iPIN (resp. iVPIN) differ from the latent PIN due to the sample variation of the estimated coefficients. The latter, due to Proposition 1 converges almost surely (as well as in the L^2 mode) to zero. Since the VPIN estimator is subject to noise due to arrival data contamination, this when correlated with the regression error, implies that the empirical MSEs of the predictions using the plug-in iPIN (resp. iVPIN) will be a.s. lower than the ones based on the interval PIN (resp. VPIN) estimator for large enough T. When the statistical model in Table 1 is misspecified, this comparison becomes a bit more complicated, as it also depends on the (non-asymptotically vanishing) dependence between the regression error, and the misspecification error provided by the difference between the latent true iPIN (resp. iVPIN) and the model iPIN (resp. iVPIN) evaluated at the pseudo-true value of the parameter.

In order however to disentangle the evaluation of the quality of the approximation of the latent PIN (resp. VPIN) by the intensity-based procedures compared to the interval-based ones, from the correct specification of a predictive regression, we perform a Monte Carlo experiment that enables control of the latent PIN. We focus on the VPIN formulations for simplicity. The volume "clock" duration process (χ_i) is assumed to conform to the specification of Table 1. The latent VPIN is thus explicitely known to the MC designer. We construct a subordinated (see [34]) to (χ_i) stochastic volatility process for the logarithmic returns of the underlying asset in the spirit of [31] as:

$$\ln P_i - \ln P_{i-1} = \alpha_0 + a_1\chi_i + \exp(\omega_0 + \omega_1\chi_i + V_i)z_i, \tag{19}$$

$$V_i = bV_{i-1} + \eta_{i-1},\ (z_i, \eta_i)^T \sim N(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}),\ |\rho| < 1, |b| < 1. \tag{20}$$

The Gaussian random vector $(z_i, \eta_i)^T$, is also considered independent of ϵ_{i-j}, for all i, j. Given initial values P_0, V_0, and for realistic sample sizes T, and choice of parameter values, artificial sample paths of (p_i, χ_i) are available. Given such a path, and using the interval-based methodology- as in Section 4.3- the interval-based VPIN estimator is producible by the researcher. For the intensity-based methodology two routes are followed: (a) the researcher correctly considers the specification of Table 1 as her underlying statistical model, performs the ML estimation and constructs the plug-in iVPIN in this context of correct specification for the duration process. (b) The researcher makes a specification error regarding ψ; she erroneously assumes that it conforms to an ACD(1,1) process instead of the correct FIACD(1,1), estimates the corresponding MLE, and constructs the plug-in iVPIN in this context of misspecification due to the presence of long memory. In all cases the estimated VPIN paths are contrasted to the latent paths of VPIN, and the Monte Carlo MSE paths between them are evaluated for comparison.

Specifically T is allowed to assume the values of 100,000 events (trades), which is the equivalent of 1 day trading in the most liquid asset in the sample, i.e., AAPL. 100 paths are considered, which is the equivalent of 100 trading days. In total, this resembles the trading activity of a very liquid assed over a calendar quarter. This, according to previous literature, is a reasonable time frame that does not introduce bias in the PIN or VPIN estimates. Furthermore, the true underlying parameter values are chosen according to the estimates in Table 3. For the price change and price change variance parameters, the following values are assumed, $\alpha_0 = -0.001, \alpha_1 = 0.9, \omega_0 = -3.5, \omega_1 = 0.05$ and $\beta = 0.98$. [13]

[13] The robustness of the findings are tested against different simulation setups and in particular in combinations of the following scenarios. The threshold value for J_i in the main scenario is 1.2. Different values that are considered are 1 and 1.9. The findings remain qualitatively the same, but are stronger when a higher threshold is employed. In addition, different values for T are considered, as well as a different number of sample paths. The number of sample paths does not change qualitatively the results, but a higher number of observations exhibits slightly less significant differences. The main scenario considers a strong negative correlation $\rho = -0.8$. Different values are tested and the results do not change significantly, but are relatively

Table 6: Simulation: Performance of metrics-Probability of Informed Trading

	1”	5”	15”	30”	1’	5’	15’	30’	60’
Average Estimates									
TRUE	0.3065	0.3384	0.3453	0.3471	0.3480	0.3487	0.3489	0.3489	0.3488
PIN	0.2659	0.2108	0.2053	0.2056	0.2059	0.2062	0.2062	0.2062	0.2062
iPIN	0.2957	0.3262	0.3326	0.3343	0.3351	0.3357	0.3359	0.3358	0.3358
VPIN	0.2372	0.2285	0.2254	0.2249	0.2248	0.2282	0.2417	0.2640	0.2933
iVPIN	0.2633	0.2904	0.2962	0.2977	0.2985	0.2991	0.2992	0.2992	0.2991
MSE									
PIN	0.0719	0.0658	0.0585	0.0576	0.0571	0.0567	0.0566	0.0566	0.0565
iPIN	0.0307	0.0288	0.0264	0.0261	0.0260	0.0259	0.0259	0.0258	0.0258
VPIN	0.1129	0.0983	0.0792	0.0779	0.0768	0.0742	0.0642	0.0518	0.0357
iVPIN	0.0207	0.0187	0.0161	0.0161	0.0157	0.0156	0.0156	0.0156	0.0156
MSE Ratios									
PIN/iPIN	2.3413	2.2857	2.2142	2.2065	2.1988	2.1909	2.1898	2.1882	2.1855
VPIN/iVPIN	5.4629	5.2443	4.9083	4.8283	4.8986	4.7669	4.1276	3.3302	2.2906
iPIN/iVPIN	1.4869	1.5347	1.6363	1.6174	1.6554	1.6617	1.6608	1.6602	1.6588
PIN/VPIN	0.6373	0.6689	0.7381	0.7391	0.7431	0.7637	0.8811	1.0909	1.5826
PIN/iVPIN	3.4813	3.5078	3.6230	3.5687	3.6401	3.6405	3.6369	3.6327	3.6253

Table 6 presents the performance of PIN, VPIN, iPIN and iVPIN in capturing the real PIN, estimated based on

Table 6 presents the results of the simulation, following the main scenario. The first line of top panel of the table presents the on-average ”true” values of the (P)robability of (IN)formed trading and it is organized in columns according to the sampling frequency that varies from 1” to 60’. The following 4 lines present the average estimates of all the metrics considered, following to empirical procedure described in Section 4.2, namely PIN, VPIN, iPIN and iVPIN. The second (middle) panel presents the MSE for each on the metrics in each sampling frequency. The bottom panel reports the ratios of MSE’s for visual reference. The findings reported here are fully in line and confirm the basic findings of the empirical analysis presented in Section 4.3. In summary, the intensity-based metrics outperform consistently the interval-based metrics in accurately capturing the ”true” PIN, by a factor of at least 2. This is consistent in all sampling frequencies, and, although it shows a decreasing trend, the intensity-based metrics are by far a better measure for private information comapared to their interval-based counterparts. The second finding that is also confirmed here is that

weaker with lower correlation.

volume does indeed contribute in capturing private information, but this is much stronger when it is associated with its aggregation properties, captured by the arrival rate of trading volume. iVPIN is consistently better than iPIN and the performance of VPIN increases considerably past the 15' sampling frequencies.

6 Conclusion

Financial markets are perceivable as the meeting place of various agent-types, who interact with each other, resolving their information sets and moving a dynamic price equilibrium ([26]). The identification of the composition of these agent-types is an essential element in describing the properties of this equilibrium, but this is a piece of latent information.

Trying to identify their existence, prior literature reaches a consensus that this might be feasible by deciphering their material actions, as they become public information in a collective manner. Each agent-type is understood to have an intrinsic motivation for trading, such as access to information (e.g., [50]) prior to competing agents (e.g., [46]), learning from it (e.g., [1]) or processing it (e.g., [57]), which is reflected on their interactions with the market. This results in variations in observable trading characteristics that might not be observable at an individual level, but, theory suggests, that their aggregated outcomes might capture the emergence properties ([5]) of their interactions and thus, they might be reverse-engineered to capture their existence. Naturally, previous approaches focus primarily on liquidity variations (e.g., [25]) and complement it with other observable variables, such as trade initiation (e.g., [21, 22, 23] or trading volume (e.g., [18, 24]), in order to identify their presence.

These approaches, albeit insightful, they are subject to a rather restrictive trade-off concerning the optimal time interval that captures the properties of the dynamic equilibrium (e.g., [3]). A longer time interval could alleviate the distortion emanating from noisy signals, such as trade initiation, while a shorter time interval would match better the needs of high frequency trading, where “to be uninformed is to be slow” ([35]). Considering that algorithmic trading operates at trading frequencies shorter than the human attention span of 650ms ([43]) a point, rather than an interval, estimate of the probability of the existence of different agent-types would be more appropriate.

This is the primary objective of this study, which ventures the idea of detecting the presence of different agent-types, not from the aggregated properties of the trading characteristics, but from the aggregation process itself; captured by their arrival rates. Motivated by relevant literature (e.g., [33, 38, 46, 48]), the actions of each agent-type is assumed to exhibit a time-invariant arrival rate. However, unlike previous studies, a general approach, with lower distributional (e.g., [33, 38]) or conditionality (e.g., [46, 47, 48] assumptions, is proposed here. The only condition required is that the (time invariant) characteristics of each agent-type should exhibit detectable patterns in any set of observable factors.

This enables the exact modelling of agent-specific arrival rates, which are then assumed to interact conditionally on market conditions. The market is seen as an infinite mixture of the agent-specific intensities and the conditional probabilities associated with each agent-specific arrival rate can be interpreted as the instantaneous probability of an event to be associated with each agent-type. Using both the instantaneous probabilities and the time-invariant agent-specific arrival rates, the probability of the presence of each agent-type can be estimated over any desirable time interval.

This shift from an interval (aggregated) to a granular (instantaneous) time frame is facilitated by the use of conditional intensities (hazard functions), and constitutes the main contribution of this paper. The detection of a multitude of various agent-types using intensity-based metrics, i.e., point estimates exhibits some notable advantages over previous studies. First, it offers an instantaneous estimate of the relevant probabilities and thus, it is interval free without suffering from sampling bias. Second, these instantaneous probabilities can be integrated over any time interval and thus, they can be used in shorter investment horizons (e.g., high-frequency trading). Third, the different agent-types are associated with a statistical measure (i.e., conditional intensity), rather than with the magnitude of a particular variable (e.g., trade or volume imbalance). Consequently, the new estimates are derived with a lower set of assumptions. Finally, the new framework can be used to detect a multitude of different agent-types, as long as their actions can be mapped into a specific shape of the hazard function and be described by a/any set of observable variables.

Overall, the framework proposed here, exhibits superior empirical and theoretical properties, and highlights the importance of time in intraday price discovery. The focus on the time dimension – how different events evolve over time – enables the development of instantaneous indicators for the presence of different, otherwise hidden, agent-types.

Time will tell!...

References

[1] Admati, A. R. and P. Pfleiderer (1986): "A monopolistic market for information," *J. Econ. Theory*, 39, 400–438.

[2] Amari, S.-i. (2016): *Information geometry and its applications*, vol. 194, Springer.

[3] Andersen, T. G. and O. Bondarenko (2014): "VPIN and the flash crash," *J. Fin. Mark.*, 17, 1–46.

[4] Andrews, D. W. (1999): "Estimation when a parameter is on a boundary," *Econometrica*, 67, 1341–1383.

[5] Arthur, W. B. (2013): "Complexity economics," *Complexity and the Economy.*

[6] Austern, M. and P. Orbanz (2022): "Limit theorems for distributions invariant under groups of transformations," *The Annals of Statistics*, 50, 1960–1991.

[7] Bacry, E., I. Mastromatteo, and J.-F. Muzy (2015): "Hawkes processes in finance," *Market Microstructure and Liquidity*, 1, 1550005.

[8] Bauwens, L. and P. Giot (2003): "Asymmetric ACD models: Introducing price information in ACD models," *Empir. Econ.*, 28, 709–731.

[9] Blasques, F., P. Gorgi, S. J. Koopman, and O. Wintenberger (2018): "Feasible invertibility conditions and maximum likelihood estimation for observation-driven models," .

[10] Bowsher, C. G. (2007): "Modelling security market events in continuous time: Intensity based, multivariate point process models," *J. Econom.*, 141, 876–912.

[11] Box, G. E. P. and D. R. Cox (1964): "An analysis of transformations," *J. R. Stat. Soc.*, 26, 211–243.

[12] Buccheri, G., F. Corsi, and S. Peluso (2021): "High-frequency lead-lag effects and cross-asset linkages: A multi-asset lagged adjustment model," *J. Bus. Econ. Stat.*, 39, 605–621.

[13] Caporin, M. (2003): "Identification of long memory in GARCH models," *Statistical Methods and Applications*, 12, 133–151.

[14] Chakrabarty, B., R. Pascual, and A. Shkilko (2015): "Evaluating trade classification algorithms: Bulk volume classification versus the tick rule and the Lee-Ready algorithm," *Journal of Financial Markets*, 25, 52–79.

[15] De Luca, G. and G. M. Gallo (2004): "Mixture Processes for Financial Intradaily Durations," *Stud. Nonlinear Dyn. Econom.*, 8.

[16] De Luca, G. and P. Zuccolotto (2006): "Regime-switching Pareto distributions for ACD models," *Comput. Stat. Data Anal.*, 51, 2179–2191.

[17] Easley, D., M. L. de Prado, and M. O'Hara (2016): "Discerning information from trade data," *J. Financ. Econ.*, 120, 269–285.

[18] Easley, D., M. L. De Prado, and M. O'Hara (2011): "The microstructure of the Flash Crash," *Journal of Portfolio Management*, 37, 118–128.

[19] Easley, D., M. M. L. de Prado, and M. O'Hara (2014): "VPIN and the flash crash: A rejoinder," *Journal of Financial Markets*, 17, 47–52.

[20] Easley, D., R. F. Engle, M. O'Hara, and L. Wu (2008): "Time-varying arrival rates of informed and uninformed trades," *Journal of Financial Econometrics*, 6, 171–207.

[21] Easley, D., N. M. Kiefer, and M. O'Hara (1997): "The information content of the trading process," *J. Empir. Finance*, 4, 159–186.

[22] ——— (1997): "One day in the life of a very common stock," *Rev. Financ. Stud.*, 10, 805–835.

[23] Easley, D., N. M. Kiefer, M. O'hara, and J. B. Paperman (1996): "Liquidity, information, and infrequently traded stocks," *J. Finance*, 51, 1405–1436.

[24] Easley, D., M. M. López de Prado, and M. O'Hara (2012): "Flow toxicity and liquidity in a high-frequency world," *Rev. Financ. Stud.*, 25, 1457–1493.

[25] Easley, D. and M. O'hara (1992): "Time and the process of security price adjustment," *J. Finance*, 47, 577–605.

[26] Easley, D. and M. O'Hara (1995): "Market microstructure," *Handbooks in operations research and management science*, 9, 357–383.

[27] Easley, D. and M. O'hara (2010): "Microstructure and ambiguity," *J. Finance*, 65, 1817–1846.

[28] Ellis, K., R. Michaely, and M. O'Hara (2000): "The accuracy of trade classification rules: Evidence from NASDAQ," *J. Fin. Quant. Anal.*, 35, 529.

[29] Engle, R. F. and J. R. Russell (1998): "Autoregressive conditional duration: A new model for irregularly spaced transaction data," *Econometrica*, 66, 1127.

[30] Fama, E. F. (1965): "The behavior of stock-market prices," *J. Bus.*, 38, 34–105.

[31] Feng, D., P. X.-K. Song, and T. S. Wirjanto (2015): "Time-deformation modeling of stock returns directed by duration processes," *Econometric Reviews*, 34, 480–511.

[32] Foucault, T., O. Kadan, and E. Kandel (2005): "Limit order book as a market for liquidity," *Rev. Financ. Stud.*, 18, 1171–1217.

[33] Gerhard, F. and N. Hautsch (2007): "A dynamic semiparametric proportional hazard model," *Stud. Nonlinear Dyn. Econom.*, 11.

[34] Ghysels, E., C. Gouriéroux, and J. Jasiak (1995): "Market time and asset price movements: Theory and estimation," *Cahier de recherche*.

[35] Haldane, A. G. (2012): "The race to zero," in *The Global Macro Economy and Finance*, London: Palgrave Macmillan UK, 245–270.

[36] Harris, L. (1989): "A day-end transaction price anomaly," *J. Fin. Quant. Anal.*, 24, 29.

[37] Hawkes, J., P. Ney, and S. Port (1976): "Advances in probability and related topics," *J. R. Stat. Soc. Ser. A*, 139, 408.

[38] Hujer, R. and S. Vuletić (2007): "Econometric analysis of financial trade processes by discrete mixture duration models," *J. Econ. Dyn. Control*, 31, 635–667.

[39] Ibragimov, I. A. and Y. V. Linnik (1971): *Independent and stationary sequences of random variables*, Wolters-Noordhoff.

[40] Jakubowski, A. (2012): "Principle of Conditioning revisited," *Demonstratio Mathematica*, 45, 325–36.

[41] Jasiak, J. (1999): “Persistence in intertrade durations,” *Finance*, 16, 166–195.

[42] Jenkinson, O. (2019): “Ergodic optimization in dynamical systems,” *Ergodic Theory and Dynamical Systems*, 39, 2593–2618.

[43] Johnson, N., G. Zhao, E. Hunsader, H. Qi, N. Johnson, J. Meng, and B. Tivnan (2013): “Abrupt rise of new machine ecology beyond human response time,” *Sci. Rep.*, 3, 2627.

[44] Johnson, N. F., P. Jefferies, and P. M. Hui (2003): *Financial market complexity*, Oxford Finance Series, London, England: Oxford University Press.

[45] Jovanovic, B. and R. W. Rosenthal (1988): “Anonymous sequential games,” *J. Math. Econ.*, 17, 77–87.

[46] Kalaitzoglou, I. and B. M. Ibrahim (2013): “Does order flow in the European Carbon Futures Market reveal information?” *J. Fin. Mark.*, 16, 604–635.

[47] Kalaitzoglou, I. A. and B. M. Ibrahim (2015): “Liquidity and resolution of uncertainty in the European carbon futures market,” *Int. Rev. Fin. Anal.*, 37, 89–102.

[48] ——— (2023): “Market conditions and order-type preference,” *Int. Rev. Fin. Anal.*, 87, 102559.

[49] Keim, D. B. and A. Madhavan (1995): “Anatomy of the trading process empirical evidence on the behavior of institutional traders,” *J. Financ. Econ.*, 37, 371–398.

[50] Kyle, A. S. (1985): “Continuous auctions and insider trading,” *Econometrica*, 53, 1315.

[51] Lee, P. M. (1989): *Bayesian statistics*, Oxford University Press London:.

[52] Lof, M. (2012): “Heterogeneity in stock prices: A STAR model with multivariate transition function,” *J. Econ. Dyn. Control*, 36, 1845–1854.

[53] MacKay, D. J. (2003): *Information theory, inference and learning algorithms*, Cambridge university press.

[54] Madhavan, A. (2000): "Market microstructure: A survey," *J. Fin. Mark.*, 3, 205–258.

[55] Nadarajah, S. and S. Kotz (2007): "On the -type distributions," *Physica A*, 377, 465–468.

[56] O'Hara, M. (2003): "Presidential address: Liquidity and price discovery," *J. Finance*, 58, 1335–1354.

[57] ——— (2015): "High frequency market microstructure," *J. Financ. Econ.*, 116, 257–270.

[58] Patterson, D. M. and V. Sharma (2010): "The incidence of informational cascades and the behavior of trade interarrival times during the stock market bubble," *Macroecon. Dyn.*, 14, 111–136.

[59] Pöppe, T., S. Moos, and D. Schiereck (2016): "The sensitivity of VPIN to the choice of trade classification algorithm," *J. Bank. Financ.*, 73, 165–181.

[60] Pyke, R. (1961): "Markov renewal processes: Definitions and preliminary properties," *Ann. Math. Stat.*, 32, 1231–1242.

[61] Ruijsenaars, S. N. (2000): "On Barnes' multiple zeta and gamma functions," *Advances in Mathematics*, 156, 107–132.

[62] Sarkar, A. and R. A. Schwartz (2009): "Market sidedness: Insights into motives for trade initiation," *J. Finance*, 64, 375–423.

[63] Shannon, C. E. (1948): "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27, 379–423.

[64] Shannon, C. E. and W. Weaver (1949): *The mathematical theory of communication, by CE Shannon (and recent contributions to the mathematical theory of communication), W. Weaver*, University of illinois Press.

[65] Straumann, D. (2006): *Estimation in conditionally heteroscedastic time series models*, vol. 181, Springer Science & Business Media.

[66] Teräsvirta, T. (1994): "Specification, estimation, and evaluation of smooth transition autoregressive models," *J. Am. Stat. Assoc.*, 89, 208–218.

[67] Tsallis, C. (1988): "Possible generalization of Boltzmann-Gibbs statistics," *J. Stat. Phys.*, 52, 479–487.

[68] Vaart, A. v. d. and J. A. Wellner (2023): "Statistical Applications," in *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer, 385–591.

Appendix A Supplementary Tables

This appendix presents more information on the descriptive statistics of the dataset employed in the empirical application of the iRP metric, as well as the estimation results for further applications of the iRP, introduced in Appendix B.2.

- Table A.1 presents the descriptive statistics for the full sample period.
- Table A.2 presents the estimation results for two further applications of the iRP introduced in sections B.2.1 and B.2.2.

The sample consists of rather liquid stocks like AAPL with average duration 0.21 sec (1.21 minus 1 sec added for computational reasons) to relatively less liquid stocks, like AXP with average duration of 1.41 sec (2.41 minus 1 sec). The average volume per trade also varies from relatively low values, like in TRV with 115.7 stocks per trade, to almost triple volume, like in PFE with 393.2 stocks per trade. In addition, price change variance exhibits a wide range of values from 0.01 in KO or VZ to 0.12 in BA covering stocks with different intensities of price discovery. Besides the relative variation that is to be expected due to the presence of cross-sectional fixed effects, no major outliers are observed, while min and max values are comparable across stocks. This implies that the sample is relatively homogeneous, but with adequate variation, in order to provide a sample with minimal trading biases or extreme events.

Table A.1: Descriptive Statistics

	D	V	R	D	V	R	D	V	R	D	V	R	D	V	R
	AAPL #25,598,759			AXP #3,836,091			BA #8,058,521			CAT #4,801,058			CSCO #10,808,593		
max	324.55	22499	4.57	2577.12	29200	1.02	840.91	12500	9.77	3211.30	28500	2.29	1439.39	53910	0.92
mean	1.21	212.10	0.00	2.41	160.58	0.00	1.66	134.46	0.00	2.14	167.88	0.00	1.48	346.35	0.00
min	1.00	1.00	-4.57	1.00	1.00	-1.02	1.00	1.00	-9.76	1.00	1.00	-2.29	1.00	1.00	-0.92
std	0.48	510.42	0.05	3.64	394.25	0.02	1.60	296.68	0.12	3.09	389.42	0.03	1.35	975.34	0.01
	CVX #6,661,469			DIS #10,296,242			DOW #3,080,034			DWDP #2,052,370			GS #3,779,536		
max	2548.29	32595	2.12	888.23	31708	1.83	4482.68	71100	1.27	154.91	78320	1.54	2543.17	19441	2.01
mean	1.81	167.86	0.00	1.52	185.19	0.00	2.31	225.44	0.00	1.70	289.63	0.00	2.44	137.43	0.00
min	1.00	1.00	-2.12	1.00	1.00	-1.83	1.00	1.00	-1.28	1.00	1.00	-1.54	1.00	1.00	-2.01
std	2.27	430.45	0.03	1.16	493.48	0.03	4.25	819.56	0.02	1.38	858.93	0.02	3.46	294.93	0.04
	HD #5,840,233			IBM #4,735,123			INTC #11,300,740			JNJ #7,406,086			JPM #10,205,497		
max	3795.37	20785	3.44	3472.45	19811	2.25	1245.15	61898	1.01	1948.77	27900	2.24	1796.65	25000	1.92
mean	1.90	128.54	0.00	2.16	145.18	0.00	1.47	363.21	0.00	1.73	166.24	0.00	1.52	218.66	0.00
min	1.00	1.00	-3.44	1.00	1.00	-2.25	1.00	1.00	-1.01	1.00	1.00	-2.24	1.00	1.00	-1.92
std	2.54	302.61	0.06	2.84	311.96	0.04	1.32	1032.44	0.01	1.99	411.58	0.04	1.42	486.70	0.03
	KO #7,288,916			MCD #4,721,261			MRK #7,413,220			MSFT #18,867,325			NKE #5,536,203		
max	1393.04	66002	0.50	2925.53	25680	2.31	2112.19	49900	1.11	762.11	27514	2.03	2478.18	49900	0.89
mean	1.74	299.68	0.00	2.14	130.48	0.00	1.71	230.84	0.00	1.29	237.32	0.00	1.99	209.79	0.00
min	1.00	1.00	-0.50	1.00	1.00	-2.31	1.00	1.00	-1.11	1.00	1.00	-2.03	1.00	1.00	-0.89
std	1.74	909.80	0.01	2.72	327.95	0.04	2.05	586.03	0.02	0.68	569.69	0.03	2.46	582.56	0.01
	PFE #10,421,158			PG #6,640,793			TRV #2,064,457			UNH #5,833,071			UTX #4,066,815		
max	1520.34	88100	0.56	2886.51	34495	1.59	2428.43	20400	1.18	4220.96	24900	5.18	4410.39	36000	3.87
mean	1.51	393.16	0.00	1.82	194.57	0.00	3.60	115.71	0.00	1.94	146.00	0.00	2.32	153.62	0.00
min	1.00	1.00	-0.56	1.00	1.00	-1.59	1.00	1.00	-1.18	1.00	1.00	-5.18	1.00	1.00	-3.87
std	1.47	1186.96	0.01	2.21	486.70	0.02	5.58	268.54	0.03	2.88	362.48	0.09	3.83	435.15	0.04
	V #7,998,274			VZ #8,121,704			WBA #4,625,492			WMT #6,213,580			XOM #7,911,163		
max	2632.40	19859	2.19	1504.48	56250	0.84	4147.27	49900	0.81	789.80	24957	1.10	1265.12	35000	1.15
mean	1.69	150.36	0.00	1.64	280.82	0.00	2.16	234.04	0.00	1.86	179.76	0.00	1.67	251.62	0.00
min	1.00	1.00	-2.20	1.00	1.00	-0.84	1.00	1.00	-0.81	1.00	1.00	-1.10	1.00	1.00	-1.16
std	1.76	323.16	0.03	1.71	786.02	0.01	3.62	601.91	0.01	1.78	411.65	0.02	1.63	588.10	0.02

Table A.1 presents the descriptive statistics, i.e., the maximum (max), the average (mean), the minimum (min) and the standard deviation (std) Duration (D in seconds; 1 second is added to all observations for computational reasons), Trading Volume (V in number of stocks) and of price change (R in $'s). The statistics are presented for the full sample period for all stocks. The numbers following the # sign is the count of observations per stock.

Table A.2: Estimation Results

	iPTT			iPTT (wating)	
	low	high		low	high
ω	0.6746		ω	0.6962	
	(0.01)			(0.01)	
β	0.4295		β	0.4183	
	(0.01)			(0.01)	
ϕ	0.4324		ϕ	0.4419	
	(0.01)			(0.01)	
δ	0.0656		δ	0.0811	
	(0.01)			(0.01)	
$(q\|E(rt))$	0.8503	1.0807	$(q\|E(rt))$	0.7242	1.4298
	(0.03)	(0.02)		(0.02)	(0.03)
$(\tau\|ti)$	1.1275	0.9522	$(\tau\|ti)_{duration<j_{duration}}$	1.0942	0.6065
	(0.01)	(0.05)		(0.01)	(0.06)
			$(\tau\|ti)_{duration<j_{duration}}$	1.6323	0.9048
				(0.02)	(0.06)
g_{ti}	0.9999		g_{ti}	0.9958	
	(0.02)			(0.02)	
			$g_{duration}$	1.0001	
				(0.02)	
$g_{E(rt)}$	0.9665		$g_{E(rt)}$	0.9950	
	(0.02)			(0.03)	
j_{ti}	1.0046		j_{ti}	1.0023	
	(0.03)			(0.05)	
			$j_{duration}$	0.9966	
				(0.04)	
$j_{E(rt)}$	1.0090		$j_{E(rt)}$	1.1052	
	(0.04)			(0.02)	

The top panel of Table A.2 presents the estimation results for the conditional mean specification parameters, assuming a FI-ACD specification $\omega + \beta\psi_{i-1} + (\chi_i - \beta\chi_{i-1}) - (\tilde{\chi}_i - \phi\tilde{\chi}_{i-1})$. The bottom panel presents the distribution parameter estimates, assuming a $q-$Weibull distribution for χ, i.e., $f\left(\chi_i|\mathcal{F}_{i-1}\right) = (2-\tau^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right)$, where $A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}}/\psi_i\right]$, $\tau_i^{q=2} = \left(G_{m=1,i}^{q=2} - G_{m=2,i}^{q=2}\right)\tau_{m=1}^{q=2} + G_{m=2,i}^{q=2}\tau_{m=2}^{q=2}$ and $G_{m=2,i}^{q=2} = \left(1+e^{-g_{m=2}^{q=2}(J_i - j_{m=2}^{q=2})}\right)^{-1}$. The collumns of Table A.2 present the estimates for the specifications of iPTT introduced in Section B.2. All estimates are cross-sectional averages, with standard deviations in (:).

Appendix B iRP and Further Applications

This appendix provides supplementary material that highlights the qualitative characteristics of iRP, as well as it discusses further potential applications.

- Section B.1 discusses how the new metric is aligned with the attributes of the marked viewed as a complex system.

- Section B.2 presents further ”indicative” applications of iRP. Section B.2.1 presents an empirical specification that can identify the presence of technical trading alongside information, while Section B.2.2 presents another customization that can identify a finer (sub-)classification of technical trading that considers waiting costs.

B.1 The market as a complex system

The formulation in Eq.(3) is also inspired and consistent with designing the market as a complex system. [44] argue that financial markets are more than "complicated" systems. They are "complex" systems, in the sense that the observable outputs of the interactions of market participants (agent-types) are more than the sum of their actions; a concept known as emergence properties. This becomes increasingly relevant at higher trading frequencies, where trading is dominated by AI-agents. These agents observe the market continuously and infer "fundamental" information from observable "trading" information, depending on their processing capacity. In parallel, their actions also generate "trading" information (information propagation) and the market is a dynamic equilibrium between information storage and propagation ([53]). [64] suggests that there is an optimal balance between information propagation and information storage capacity that maximizes reward per unit of effort, a concept known as maximum informational entropy. [5] argues that this optimal level is an emergence property, which should be modelled alongside the actions of individual agents.

This specific point is a major contribution over previous approaches of Eq.(3), which indirectly addresses all the properties of a complex system (e.g., [44]):

Multiple interacting agents: Eq.(3) proposes a market design where multiple agents with intensities $\lambda_0^k(t)$ interact and their actions collectively create the market wide intensity $\lambda^k(t|\mathcal{F}_s)$. Consequently, the overall market activity is the output of the actions of all agents present in the market, $\lambda_0^k(t)$, as well as their interactions, $(p_t^k|\mathcal{F}_s)$, which are conditional on the market as a whole. This is because the information set that determines the overall market activity does so through the conditional probabilities of each agent entering the market.

Adaptation: Adaptation refers to the ability of individual agents to adjust their behaviour in order to improve their performance. This is directly modelled in Eq.(3) with the weighting probabilities, $(p_t^k|\mathcal{F}_s)$. $(p_t^k|\mathcal{F}_s)$'s are conditional on market conditions, $(:|\mathcal{F}_s)$, and define the probability of a specific agent-type with intensity $\lambda_0^k(t)$ to enter the market. This is done in a manner conditional on market activity and not independently of it and, therefore, different agent-types can adapt their behaviour, i.e., adjust the weight of $\lambda_0^k(t)$, according to market conditions, $(: |\mathcal{F}_s)$, which include the actions of all agents.

Feedback: The adaptive learning described above is designed in way that also considers information propagation. $(p_t^k|\mathcal{F}_s)$'s are conditional on observable information, which according to Eq.(3) is necessarily the collective output of all interacting agents. This way, Eq.(3) establishes a feedback mechanism that considers learning and information processing through the conditional set, $(: F_s)$, as well as through the constraint that overall market activity, $\lambda^k(t|\mathcal{F}_s)$, is necessarily the output of individual agent actions, $\lambda_0^k(t)$, i.e., $\sum_{k=1}^K (p_t^k|\mathcal{F}_s) = 1$.

Evolution: The information set $(: |\mathcal{F}_s)$ does not necessarily contain only endogenous information, i.e., $\lambda^k(t|\mathcal{F}_s)$, but it can also contain other exogenous variables. This way, exogenous, as well as endogenous, shocks can affect the probabilities of different agent-types to enter the market, $(p_t^k|\mathcal{F}_s)$, without necessarily imposing a mean reversion property.[14] Consequently, the market activity, $\lambda(t|\mathcal{F}_s)$, can evolve, continuously searching for an equilibrium, and can even exhibit 'extreme behaviour', such as crashes or bubbles. This would happen when $\lambda^k(t|\mathcal{F}_s)$ takes extreme values, which would then be also fed back to the system through $(: |\mathcal{F}_s)$'s, potentially leading to market failures.

Non-stationarity: This evolution does not have to be stationary, i.e., system properties observed in the past do not necessarily remain unchanged in the future. Eq.(3) operates at a trade-off with respect to stationarity. The agent-specific characteristics $\lambda_0^k(t)$, i.e., the way agent-types act given a specific attribute, such as learning, speed, processing capacity etc., are assumed to be stationary. This is done for traceability reasons and not because it represents better complex system properties. This is an explicit assumption, which implies that an agent-type reacts to the same stimuli in a way that does not change over time. According to the formulation in Eq.(3), this is what makes her actions distinguishable and

[14] [7] provides a martingale representation of Hawke's processes, while [29] suggest a way to model the the innovations of inter-event waiting times, i.e., durations. These two are the two most popular approaches employed in finance to model time and both have an innovation component embedded. This way endogenous or exogenous shocks can affect the arrival rate of either agent-specific or market-wide events. In this paper, due to its empirical focus, the approach of [29] is preferred due to the explicit modelling of the innovations, which will be the main tool to distinguish among different agent-types.

thus, traceable. This is definitely a constrain in the modelling, but Eq.(3) can accommodate time variant agent-type characteristics, albeit in a rigid way. Assuming that the information processing capacity of an agent-type changes due to a structural break, such as changes in technology. This would necessarily imply a change of the baseline intensity, $\lambda_0^{kA}(t) \rightarrow \lambda_0^{kB}(t)$. Eq.(3) would be able to identify $\lambda_0^{kB}(t)$ (i.e., as a new intensity) and also $(p_t^{kB}|\mathcal{F}_s)$, making at the same time $(p_t^{kA}|\mathcal{F}_s) = 0$.

B.2 Further Applications: Beyond Information

The formulation of iRP in Eq.(12) is rather generic and it can cover a considerably wider range of different agent-types beyond referring solely to information and information diffusion. The underlying concept is that if the actions of a specific agent-type are reflected on some observable factors of market activity (captured by the threshold variable(s), i.e., $\boldsymbol{J}_i$), which in turn have an impact on an observable trading pattern (the impact of $\boldsymbol{J}_i$ on the arrival rate, i.e., the hazard function, i.e., $\lambda^k(t)$ of specific variable, i.e., the modelled variable, i.e., χ_i); then Eq.(12) can estimate the instantaneous probability of their existence. In order to "reverse engineer" the probability of a specific agent-type to enter the market from the observed trading activity, the building blocks of Eq.(12) are:

Economic Concept	**Statistical Variable**	**Notation**
Observable Factors	Set of threshold variables	$\boldsymbol{J}_i$
The speed of events	The modelled variable	χ_i
The arrival rate of the events	The Hazard Function	$\lambda^k(t)$
Variations in observable factors	Regimes of the threshold variables	$J_i^q < \jmath_{m=2}^q$ ... $\jmath_{m=\mathcal{M}-1}^q \leq J_i^q < \jmath_{m=\mathcal{M}}^q$ ...$J_i^q \geq \jmath_{m=M}^q$
Variations in the arrival rates	Different distributions defined by shape/scale parameters	τ_m^q

This section presents an application of the iRP through various extensions of the framework presented in Eqq.(4)-(11). More precisely, Section 3.2 employs a different modelled variable, i.e., the speed of volume accumulation, in order to capture order flow toxicity, similar to VPIN. Section B.2.1 expands the interpretation of iPIN by dissecting the shape and the scale parameter simultaneously in order to investigate technical trading. Finally, Section B.2.2 dissects the shape parameter further, i.e., in two dimensions, in order to investigate the behavioral bias of ”patient traders” (e.g., [32]). In brief, the following sections employ different specifications of Eqq.(4)-(11) and link them indicatively to different agent-types, as an illustration of the flexibility of the framework.

B.2.1 Information and Market-Wide Learning: Technical Trading

Access to private information might not be synchronous to all traders and, therefore, the classification of informed versus uninformed (e.g., [50]) might be unrealistic. Previous literature recognizes the existences of other groups under the general notion of discretionary liquidity traders (e.g., [1]), usually referred to as “technical” traders. Although there are many sub-classifications according to their way/speed of learning or access to information (e.g., [57]), they tend to exhibit some common characteristics. They are generally understood as a challenge to weak-form market efficiency. They observe the market and “learn”, in the sense of inferring exploitable patterns from trading history. They only enter the market when they (believe they) extract price-relevant information, in a discretionary way by selecting the timing and the volume of their trading. Unlike the uninformed agents, whose arrival rate is time invariant, these partially uninformed agents try to become better informed (learn). Consequently, their arrival rate depends on market conditions.

[46] suggest that technical traders, due to their learning time-related cost, react to information with a delay compared to better informed agents. Consequently, their probability of entering the market, following an information signal increases with time, resulting in a hazard function with an increasing shape. Other studies, however, focus on market signals that might instigate technical trading, with expected return being probably the strongest indicator. For example, [8] imply that the trading activity of technical traders increases after large price changes are expected and therefore, their probability to enter the market is higher closer to the events that result in higher expected price changes and, then, it decreases over

time. This implies that when $J_i := \mathbb{E}(R_i|\mathcal{F}_{i-1}) > j_{\text{technical}}$, then the hazard function should exhibit a decreasing shape. Both approaches imply a monotonic hazard function and are constrained by the use of the distributions employed, as well as the observable threshold variables.

Eq.(12) provides a versatile tool for identifying technical trading, using multiple observable factors. Combining both approaches, technical traders are understood to observe past price changes and formulate expectations about the presence of information signals that motivate their discretionary interaction with the market. Naturally, they should be expected to act as soon as they can, before the informational advantage expires by becoming public information. This should result in a hazard function with a decreasing shape. However, they cannot enter the market immediately after the arrival of new information because they do not possess it from the beginning. They have to extract it first and, therefore, their probability of entering the market after the arrival of information should increase over time, resulting in a hazard function with increasing shape. Consequently, the trading behaviour of technical traders should be a combination of the two different shapes of the hazard function – first increasing (enter the market with a delay) and then decreasing (act timely on newly acquired information) until it reaches zero when information becomes public – implying a unimodal shape.[15]

Eq.(13) is flexible enough to capture non-monotonic hazard functions and the one that matches the unimodal shape is when $\tau^{q=1} \geq 1$ and $\tau^{q=2} \geq 1$. Following, the formulation of iPIN and iVPIN, the actions of technical traders are assumed to be expressed primarily in trading intensity, prior to any significant price change that would incorporate the new information. They act on information with a delay, because they have to "learn" first, and there-

[15] In the absence of information, the probability of technical traders should be low or zero. In the presence of ”extracted” information they should act within its ”life span”, implying a decreasing hazard function. However, because they have to ”learn” first, they cannot act as fast as the informed agents and their actions are expressed with a delay. This implies an increasing hazard function upon the arrival of information. When inforamtion is price-resolved there is no monetary benefit and, therefore, their probability of entering the market returns to zero. Collectively, the probability of technical traders follows a unimodal shape that cannot be captured by single-shape parameter distributions.

fore, their hazard function should exhibit mostly an increasing shape, at least during the "life span" of the information. In line with the formulation of of iPIN and iVPIN, trading intensity is assumed to be a primary characteristic that expresses the actions of different agent-types and therefore, it is also used here as a major determinant of the shape of the hazard function, i.e., flat/increasing/decreasing. Following the notation in Eq. (13) and the example in Table 1 the shape parameter ,i.e., $\left(\tau_i^{q=2}\middle|J_{v=1,i}^{q=2}=ti_i\right)$, is conditional on trading intensity. Considering that a decreasing hazard function is associated with informed trading (e.g., iPIN), a milder or increasing (e.g., [46]) shape of the hazard function should be associated with less informed trading that is more consistent with the presence of technical traders. This on its own is not sufficient and a secondary signal is also considered. The magnitude of the reaction of technical traders is proportional to the magnitude of the inferred information and this is in lined with the impact (extensivity) of the entropy parameter on the shape of the hazard function. The "informativeness" of the magnitude of an observation and its subsequent impact on the shape of the hazard function is captured by the entropy parameter, $\tau_i^{q=1}$, which is assumed to be conditional on the magnitude of the (information) signal, here assumed to be only (but this is only indicative) expected returns, i.e., $\left(\tau_i^{q=1}\middle|J_{v=1,i}^{q=1}=\mathbb{E}(R_i|\mathcal{F}_{i-1})\right)$. Higher values should be associated with higher presence of technical trading. It follows that the intensity-based probability of technical traders, i.e., iTT, can be defined following Eq.(12) as: iTT $=\left(\mathrm{iRP}_t^{\mathrm{technical}}\middle|\mathcal{F}_s\right)=\frac{\mathbb{E}\left(\int_s^t\lambda^{Z=\mathrm{technical}}(u)\,du\middle|\mathcal{F}_s\right)}{\sum_{k=1}^K\mathbb{E}\left(\int_s^t\lambda^k(u)\,du\middle|\mathcal{F}_s\right)}=\frac{(p_t^{Z=\mathrm{technical}}|\mathcal{F}_s)H^{Z=\mathrm{technical}}(t|\mathcal{F}_s}{\sum_{k=1}^K((p_t^k|\mathcal{F}_s)H^k(t|\mathcal{F}_s)}$, which can then be expressed in terms of the regimes of shape parameters, τ_i^q, in Eq.(13) as:

$$\mathrm{iTT}_i=\frac{\sum_{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)}\left\{\prod_{Q\bigotimes M}W_{m,i}^qH^{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)}(t)\right\}}{\sum_{Q\bigotimes M}\prod_{Q\bigotimes M}W_{m,i}^qH^{Q\bigotimes M}(t)} \tag{B.1}$$

In this formulation, the regimes, m of $\tau^{q=1}$ and $\tau^{q=2}$ that lead to a unimodal shape of the hazard function, identify technical trading (k = technical). Then, the aggregated number of technical traders is the sum of all the probabilities, i.e., $\prod_{Q\bigotimes M}W_{m,i}^q$, of inter-

$$\psi_i = \omega + \beta\psi_{i-1} + (\chi_i - \beta\chi_{i-1}) - (\tilde{\chi}_i - \phi\tilde{\chi}_{i-1}), \text{ where } \chi_i = \Delta t_i^*$$

$$f\left(\chi_i|\mathcal{F}_{i-1}\right) = (2 - \tau_i^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right), \text{ where } A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}} / \psi_i\right]$$

$$\tau_i^{q=1} = \left(1 - G_{m=2,i}^{q=1}\right)\tau_{m=1}^{q=1} + G_{m=2,i}^{q=1}\tau_{m=2}^{q=1} \text{ and } \tau_i^{q=2} = \left(1 - G_{m=2,i}^{q=2}\right)\tau_{m=1}^{q=2} + G_{m=2,i}^{q=2}\tau_{m=2}^{q=2}$$

$$\text{for} \quad G_{m=2,i}^{q=1} = \left(1 + e^{-g_{m=2}^{q=2}(E(rt)_i - j_{m=2}^{q=2})}\right)^{-1}, \text{ and } G_{m=2,i}^{q=2} = \left(1 + e^{-g_{m=2}^{q=2}(ti_i - j_{m=2}^{q=2})}\right)^{-1}$$

sections $Q \bigotimes \left(m\left(1 < \tau_m^{q=1} \leq 2, \tau_m^{q=2} > 1\right) : k\right)$ in the contingency table, where $1 < \tau_m^{q=1} \leq 2$ and $\tau_m^{q=2} > 1$, times the respective cumulative hazard functions of these intersections, i.e., $H^{Q \bigotimes \left(m\left(1<\tau_m^{q=1}\leq 2, \tau_m^{q=2}>1\right):k\right)}(t)$. This is then compared to the expected number of all trades. Technical trading in Eq.(B.1) can be estimated in parallel with iPIN. Different regimes of $\tau_m^{q=1}$ and $\tau_m^{q=1}$ that lead to a decreasing hazard function would identify informed trading. The regimes that exhibit a unimodal shape would indicate the presence of technical trading.

Table A.2 reports the cross-sectional estimates of the specification in this table that intends to identify technical trading. In particular, the interest lies in the middle panel under iPTT, which reports the estimates for the distribution of χ with both the scale and the shape parameter being an infinite mixture of two regimes. Focusing on the shape parameter $\tau_i^{q=2}$, noted as τ_{ti}, it takes values higher than one when $ti_i < j_{ti}$ (noted as $ti < j(ti)$, while it takes values less than 1 when $ti_i > j_{ti}$ (noted as $ti > j(ti)$. In consistence with the iPIN and iVPIN estimations higher trading intensity is associated with higher presence of private information (decreasing hazard function). However, lower trading intensity that leads to $\tau_{ti} > 1$ and higher $E(rt)$ that leads to $\tau_i^{q=1} > 1$ (noted as $q_{E(rt)>j(E(rt))}$) make the hazard function take a unimodal shape. This matches the characteristics of technical trading. At the bottom panel, their cross-sectional average over the sample period is estimated around 30%.

The remaining regimes that exhibit differently shaped hazard functions could be associated with different agent-types. As an indications, previous literature (e.g., [46]) suggest that a flat hazard function can be associated with uninformed agents. Of course, there are different shapes and/or different combinations. The next section (Section B.2.2) discusses indicatively a variation of the potential shapes the hazard function can take.

B.2.2 Information Diffusion and Learning: Waiting Costs

The discussion above is based on market-wide learning and uniform access to information. However, technical traders are a diverse collection of agents (e.g., [57]) motivated by a variety of factors, such as investment style and behavioural biases. These differences would instigate different trading styles that thus, different accumulation rates (hazard functions). A potential (indicative) factor is waiting costs (e.g., [32]). Lower-frequency traders ("patient") trade according to their portfolio-re-balancing needs (e.g., [49]). They are more likely to wait until a sufficiently strong signal or submit a limit order trying to secure a better price (e.g., [32]); actions that lead to delayed execution. In contrast, faster, e.g., algorithmic, traders ("impatient") profit from accessing and acting on information faster than the remaining uninformed agents. They are more likely to submit a market order as soon as it is profitable. Both types can be considered uninformed and their actions are associated with a unimodal hazard function. However, (im-)patient traders face a (higher) lower waiting cost and therefore, the degree of the curvature of the hazard function should be (sharper) milder (e.g., [32]), even monotonically (decreasing) increasing in the limit.

An extension of Eq.(B.1) can capture this dissection of technical traders. Consider the setup, in Eq.(B.1), i.e., $\left(\tau_i^{q=2}\middle|J_{v=2,i}^{q=2}=\frac{\text{volume}_i}{\text{duration}_i}\right)$ and $\left(\tau_i^{q=1}\middle|J_{v=1,i}^{q=1}=\mathbb{E}(R_i|\mathcal{F}_{i-1})\right)$. Impatient traders act faster than patient traders and this can be captured by distinguishing the trades that result into $\tau^{q=2}>1$, i.e., technical trading, into relatively high or low duration. To capture this, $\tau^{q=2}$ is dissected into levels of trading intensity, i.e., $J_{v=2,i}^{q=2}$, as well as of duration, i.e., $J_{v=3,i}^{q=2}=\text{duration}_i$. This conditions the shape parameter on two variables, $\left(\tau_i^{q=2}\middle|J_{v=2,i}^{q=2},J_{v=3,i}^{q=2}\right)$, and dissects it into four cases; high (low) trading intensity and long (short) durations.

$$\tau_i^{q=2} = \Bigg[\underbrace{\left(1 - G^{q=2}_{v=2,m=2,i}\left(J^{q=2}_{v=2,i}\right)\right)\left(1 - G^{q=2}_{v=3,m=2,i}\left(J^{q=2}_{v=3,i}\right)\right)}_{W_{\text{low } J^{q=2}_{v=2,i}, \text{ long } J^{q=2}_{v=3,i}, i}} \tau^{q=2}_{\text{low } J^{q=2}_{v=2,i}, \text{ long } J^{q=2}_{v=3,i}}$$

$$+ \underbrace{G^{q=2}_{v=2,m=2,i}\left(J^{q=2}_{v=2,i}\right)\left(1 - G^{q=2}_{v=3,m=2,i}\left(J^{q=2}_{v=3,i}\right)\right)}_{W_{\text{high } J^{q=2}_{v=2,i}, \text{ long } J^{q=2}_{v=3,i}, i}} \tau^{q=2}_{\text{high } J^{q=2}_{v=2,i}, \text{ long } J^{q=2}_{v=3,i}} \Bigg]$$

$$\Bigg[\underbrace{\left(1 - G^{q=2}_{v=2,m=2,i}\left(J^{q=2}_{v=2,i}\right)\right) G^{q=2}_{v=3,m=2,i}\left(J^{q=2}_{v=3,i}\right)}_{W_{\text{low } J^{q=2}_{v=2,i}, \text{ short } J^{q=2}_{v=3,i}, i}} \tau^{q=2}_{\text{low } J^{q=2}_{v=2,i}, \text{ short } J^{q=2}_{v=3,i}}$$

$$+ \underbrace{G^{q=2}_{v=2,m=2,i}\left(J^{q=2}_{v=2,i}\right) G^{q=2}_{v=3,m=2,i}\left(J^{q=2}_{v=3,i}\right)}_{W_{\text{high } J^{q=2}_{v=2,i}, \text{ short } J^{q=2}_{v=3,i}, i}} \tau^{q=2}_{\text{high } J^{q=2}_{v=2,i}, \text{ short } J^{q=2}_{v=3,i}} \Bigg] \tag{B.2}$$

The formulation above dissects the shape parameter, $\tau_i^{q=2}$, into four regimes. Different combinations of trading intensity and duration will result in different levels of $\tau_i^{q=2}$ and, consequently, in different shapes of the hazard function. The shapes of interest are the variations of a unimodal shape, i.e., $\tau_i^{q=2} > 1$, that might vary from a marginally decreasing shape to a marginally increasing. In particular, when expected return is high, $\tau_i^{q=1}$ is expected to be $1 < \tau_i^{q=1} \leq 2$. In combination with a $\tau_i^{q=2} > 1$, this would lead to a unimodal shape. The identification that is pursued here goes one step further and distinguishes different levels of $\tau_i^{q=2}$ according to different levels of duration. In particular, (shorter) longer duration should be associated with (im-)patient trading, resulting in a hazard function that exhibits a (sharper, i.e., $\tau_i^{q=2} \xrightarrow{+} 1$) milder, i.e., $\tau_i^{q=2} > 1$, decrease, which in the limit could even reach a (decreasing) increasing shape. This identifies impatient technical traders (iITT) and patient technical traders (iPTT) (see e.g., [32]), with a relative proportion that can be defined as:

$$\begin{aligned}
\mathrm{iITT}_i &= \frac{\sum_{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}\xrightarrow{+}1\right):k\right)}\left\{\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}\xrightarrow{+}1\right):k\right)}(t)\right\}}{\sum_{Q\bigotimes M}\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes M}(t)} \\
\mathrm{iPTT}_i &= \frac{\sum_{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)}\left\{\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)}(t)\right\}}{\sum_{Q\bigotimes M}\prod_{Q\bigotimes M} W_{m,i}^{q} H^{Q\bigotimes M}(t)}
\end{aligned} \tag{B.3}$$

The regimes, m of $\tau^{q=1}$ and $\tau^{q=2}$ that lead to a unimodal shape of the hazard function identify the technical trading, while the different degrees of curvature are associated with the magnitude of waiting costs. The aggregated number of impatient (or patient in (:)) technical traders, is the sum of all the probabilities, i.e., $\prod_{Q\bigotimes M} W_{m,i}^{q}$, of intersections $Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}\xrightarrow{+}1\right):k\right)$ (or $Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)$ for patient traders) in the contingency table, where $1<\tau_m^{q=1}\leq 2$ and $\tau_m^{q=2}\xrightarrow{+}1$ (or $1<\tau_m^{q=1}\leq 2$ and $\tau_m^{q=2}>1$, times the respective cumulative hazard functions, $H^{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}\xrightarrow{+}1\right):k\right)}(t)$ (or $H^{Q\bigotimes\left(m\left(1<\tau_m^{q=1}\leq 2,\tau_m^{q=2}>1\right):k\right)}(t)$). This is then compared to the expected number of all trades.

As a final example, the specification below follows the one in Section B.2.1 but with a higher refinement of the shape parameter, which is now split across an additional dimension. More specifically, consider the following model:

In this specification, the entropy parameter, $q=\tau_i^{q=1}$, is split into two regimes, $m=1,2$, defined by the threshold variable, $E(rt)$, as it is compared to the threshold, $j_{m=2}^{q=1}$. The shape parameter though is split into a $2x2$ dimension matrix, defined by two variables, $J_{v=2,i}^{q=2}=\mathrm{duration}_{i-1}$ and $J_{v=3,i}^{q=2}=ti_i$. The first one captures the market conditions that might be associated with technical trading (trading intensity associated with an increasing hazard function), while the second distinguishes this level of trading intensity into faster-

$$\psi_i = \omega + \beta\psi_{i-1} + (\chi_i - \beta\chi_{i-1}) - (\tilde{\chi}_i - \phi\tilde{\chi}_{i-1}), \text{ where } \chi_i = \Delta t_i^*$$

$$f\left(\chi_i|\mathcal{F}_{i-1}\right) = (2-\tau_i^{q=1})\frac{\tau_i^{q=2}}{\chi_i}\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}} e_q\left(-\left[\frac{\chi_i}{A_i}\right]^{\tau_i^{q=2}}\right), \text{ where } A_i = \left[\Gamma\left(1+\frac{1}{\tau_i^{q=2}}\right)^{-\tau_i^{q=2}} / \psi_i\right]$$

$$\tau_i^{q=1} = \left(1 - G_{m=2,i}^{q=1}\right)\tau_{m=1}^{q=1} + G_{m=2,i}^{q=1}\tau_{m=2}^{q=1}$$

$$\tau_i^{q=2} = \left(1 - G_{v=2,i}^{q=2,a}\right)\left(1 - G_{v=3,i}^{q=2,b}\right)\tau_{J_{v=2,m=1,i}^{q=2}J_{v=3,m=1,i}^{q=2}}^{q=2} + G_{v=2,i}^{q=2,a}G_{v=3,i}^{q=2,b}\tau_{J_{v=2,m=2,i}^{q=2},J_{v=3,m=2,i}^{q=2}}^{q=2}$$

$$+G_{v=2,i}^{q=2,a}\left(1 - G_{v=3,i}^{q=2,b}\right)\tau_{J_{v=2,i}^{q=2,a},J_{v=3,m=1,i}^{q=2}}^{q=2} + \left(1 - G_{v=2,i}^{q=2,b}\right)G_{v=3,i}^{q=2,a}\tau_{J_{v=2,m=1,i}^{q=2},J_{v=3,m=2,i}^{q=2}}^{q=2}$$

$$\text{for} \quad G_{m=2,i}^{q=1} = \left(1 + e^{-g_{m=2}^{q=2}(E(rt)_i - j_{m=2}^{q=2})}\right)^{-1},$$

$$G_{m=2,i}^{q=2,a} = \left(1 + e^{-g_{v=2,m=2}^{q=2}(ti_{i-1} - j_{v=2,m=2}^{q=2})}\right)^{-1} \text{ and } G_{m=2,i}^{q=2,b} = \left(1 + e^{-g_{v=3,m=2}^{q=2}(\text{duration}_i - j_{v=3,m=2}^{q=2})}\right)^{-1}$$

bigger (shorter duration-higher volume) and slower-smaller (longer duration-lower volume) trades, which is intended to capture the execution strategy (fast or slow). Each threshold variable has its own threshold, $j_{v=2,m=2}^{q=2}$ for $J_{v=2,i}^{q=2}$ and . The magnitude of the threshold variables, $J_{v=2,i}^{q=2} = ti_{i-1}$ and $J_{v=3,i}^{q=2} = ti_i$, relative to their thresholds, $j_{v=2,m=2}^{q=2}$ and $j_{v=3,m=2}^{q=2}$, define four different combinations, with the shape of the hazard functions being defined by the related shape parameters:

	$J_{v=3,i}^{q=2} \leq j_{v=3,m=2}^{q=2}$	$J_{v=3,i}^{q=2} > j_{v=3,m=2}^{q=2}$
$J_{v=2,i}^{q=2} \leq j_{v=2,m=2}^{q=2}$	$\tau_{J_{v=2,m=1,i}^{q=2}J_{v=3,m=1,i}^{q=2}}^{q=2}$	$\tau_{J_{v=2,m=1,i}^{q=2},J_{v=3,m=2,i}^{q=2}}^{q=2}$
$J_{v=2,i}^{q=2} > j_{v=2,m=2}^{q=2}$	$\tau_{J_{v=2,i}^{q=2,a},J_{v=3,m=1,i}^{q=2}}^{q=2}$	$\tau_{J_{v=2,m=2,i}^{q=2},J_{v=3,m=2,i}^{q=2}}^{q=2}$

These parameters have been indicatively been estimated in the left side of Table A.2, in the right column under iPTT. The middle part of this column reports the estimates of the parameters of the distribution, which are inline with expectations and previous estimations. More precisely, in accordance with Section B.2.1, the values of $\tau^{q=2}$ are below 1 when $J_{v=2,m=2,i}^{q=2} > j_{v=2,m2}$ (noted as $ti > j(ti)$), which indicates informed trading. However, the shape parameter, $\tau^{q=2}$ takes values that are consistently higher than 1 when $J_{v=2,m=2,i}^{q=2} \leq j_{v=2,m2}$ (noted as $ti < j(ti)$), which in combination with a scale parameter, $\tau^{q=1} = q$, that takes values higher than 1 when $E(rt) > j_{v=3,m2}$ (noted as $E(rt) > j(E(rt))$), imply a unimodal shape for the hazard function. This is consistent with technical trading. On a deeper dissection, when duration is low (shaded areas) the shape parameter takes a value closer to 1, while longer durations are associated with higher values. This is consistent with impatient and patient technical trading. These quantities can be then estimated following Eq. (B.3) and they are reported in the bottom panel and take values aroung 15%.

Appendix C Technical Appendix

C.1 Proofs

Proof of Theorem 1. The condition $\gamma^T|p^\star - p| \to 0$ ensures the exponentially fast almost sure approximation of the truncated at $p^\star$ filter $\Psi^\star(\theta; \psi_{i-l}, x_{i-j}, l = 1, \ldots, \max(p,q),\ j = 1, \ldots, p)$ by $\Psi^\star(\theta; \psi_{i-l}, x_{i-j}, l = 1, \ldots, \max(p,q),\ j = 1, \ldots, p)$ uniformly over Θ. The proof is then identical to the proof of Theorem 4.3 of [9]. □

Proof of Proposition 1. Follows from the CMT, Corollary 1, and Proposition 3.2 of [9]. □

Proof of Theorem 2. As in Theorem 1, the condition $\gamma^T|p^\star - p| \to 0$ ensures the exponentially fast almost sure approximation of the first and second derivatives of the truncated at $p^\star$ filter $\Psi^\star(\theta; \psi_{i-l}, x_{i-j}, l = 1, \ldots, \max(p,q),\ j = 1, \ldots, p)$ by the analogous derivatives of $\Psi^\star(\theta; \psi_{i-l}, x_{i-j}, l = 1, \ldots, \max(p,q),\ j = 1, \ldots, p)$ locally uniformly over Θ. Notice that irrespective of the pseudo-true value of q, the first derivative of the likelihood contributions evaluated at $\phi^\star$ conditionally on the filtration, lies in the normal domain of attraction of a zero mean Gaussian distribution (see Theorem 2.6.5 in Ibragimov and Linnik [39]). Then stationarity and ergodicity and second order integrability of the limiting filter of the first derivatives as well as the almost sure boundedness of the derivatives of the remaining processes along with the principle of conditioning (see [40]), implies $O_p(\sqrt{T})$ asymptotic tightness and limiting zero mean Gaussianity for the score. The locaally uniform version of the ergodic theorem takes care the relevant a.s. convergence of the Hessian, and the result follows from a Mean Value expansion of the f.o.c.s. of the optimization problem that defines the estimator, which hold w.h.p. due to **B1**. □

C.2 Convergence of Alternative Agent-Types

The CMT and the Delta method, along with the limit theory derivations for the estimated parameters of interest, directly imply analogous results to Propositions 1, and 4 for the iTT and iPTT. Those are reported below for completeness of exposition:

Proposition 3 (C.2.1)**.** *Suppose that (i). assumptions **A1**-**A5** hold and* $\epsilon_T \to 0, \mathbb{P}$ *a.s., (ii). each element of* τ_i *is almost surely continuous in* ϕ*, (iii). the cumulative hazard functions employed are continuous functions of the shape and scale parameters, (iv)* $\mathbb{E}(\log^+ \sup_\theta f_0(\theta)) < +\infty$ *and (v). the statistical model is well specified. Then as* $s+1 \leq T \to \infty$, $|\mathrm{iTT}_{t,s}(\phi_T, \hat{\psi}_{s+1}) - \mathrm{iTT}_{t,s}(\phi_0)| + |\mathrm{iPTT}_{t,s}(\phi_T, \hat{\psi}_{s+1}) - \mathrm{iPTT}_{t,s}(\phi_0)| \to 0$, $\mathbb{P}$ *a.s. conditionally on* $\mathcal{F}_s$*.*

Proposition 4 (C.2.2)**.** *Under the premises of Theorem 2 and if the cumulative hazard functions employed are continuous functions of the shape and scale parameters, and the statistical model is well specified, then as* $s+1 \leq T \to \infty$*, we have for the* iTT *case that* $d_W(\sqrt{T}(\mathrm{iTT}_{t,s}(\phi_T, \hat{\psi}_{s+1}(\phi_T)) - \mathrm{iTT}_{t,s}(\phi_0)), N(0, \partial_\phi \mathrm{iTT}(\phi_0)^T (\partial_\phi \ell_0(\phi_0) \partial_\phi \ell_0^T(\phi_0))^{-1} \partial_\phi \mathrm{iTT}_{t,s}(\phi_0))) \to 0$, $\mathbb{P}$ *a.s. conditionally on* $\mathcal{F}_s$*, and furthermore likewise for the* iPTT *case we have that,* $d_W(\sqrt{T}(\mathrm{iPTT}_{t,s}(\phi_T, \hat{\psi}_{s+1}(\phi_T)) - \mathrm{iPTT}_{t,s}(\phi_0)), N(0, \partial_\phi \mathrm{iPTT}(\phi_0)^T (\partial_\phi \ell_0(\phi_0) \partial_\phi \ell_0^T(\phi_0))^{-1} \partial_\phi \mathrm{iPTT}_{t,s}(\phi_0))) \to 0$, $\mathbb{P}$ *a.s. conditionally on* $\mathcal{F}_s$*.*

The proofs are essentially the same to the ones of Propositions 1, and 4 respectively. The continuity of the hazard functions assumed holds for the specification used above. The first proposition implies consistency for both the iTT and the iPTT under correct specification. The second implies standard rates and asymptotic normality that could be useful for statistical inference.

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
SCHOOL OF ECONOMIC SCIENCES
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF ECONOMICS

Department of Economics
Athens University of Economics and Business

List of Recent Working Papers

2022

01-22 **Is Ireland the most intangible intensive economy in Europe? A growth accounting perspective, Ilias Kostarakos, KieranMcQuinn and Petros Varthalitis**

02-22 **Common bank supervision and profitability convergence in the EU, Ioanna Avgeri, Yiannis Dendramis and Helen Louri**

03-22 **Missing Values in Panel Data Unit Root Tests, Yiannis Karavias, Elias Tzavalis and Haotian Zhang**

04-22 **Ordering Arbitrage Portfolios and Finding Arbitrage Opportunities, Stelios Arvanitis and Thierry Post**

05-22 **Concentration Inequalities for Kernel Density Estimators under Uniform Mixing, Stelios Arvanitis**

06-22 **Public Sector Corruption and the Valuation of Systemically Important Banks, Georgios Bertsatos, Spyros Pagratis, Plutarchos Sakellaris**

07-22 **Finance or Demand: What drives the Responses of Young and Small Firms to Financial Crises? Stelios Giannoulakis and Plutarchos Sakellaris**

08-22 **Production function estimation controlling for endogenous productivity disruptions, Plutarchos Sakellaris and Dimitris Zaverdas**

09-22 **A panel bounds testing procedure, Georgios Bertsatos, Plutarchos Sakellaris, Mike G. Tsionas**

10-22 **Social policy gone bad educationally: Unintended peer effects from transferred students, Christos Genakos and Eleni Kyrkopoulou**

11-22 **Inconsistency for the Gaussian QMLE in GARCH-type models with infinite variance, Stelios Arvanitis and Alexandros Louka**

12-22 **Time to question the wisdom of active monetary policies, George C. Bitros**

13-22 **Investors' Behavior in Cryptocurrency Market, Stelios Arvanitis, Nikolas Topaloglou and Georgios Tsomidis**

14-22 **On the asking price for selling Chelsea FC, Georgios Bertsatos and Gerassimos Sapountzoglou**

15-22 **Hysteresis, Financial Frictions and Monetary Policy, Konstantinos Giakas**

16-22 **Delay in Childbearing and the Evolution of Fertility Rates, Evangelos Dioikitopoulos and Dimitrios Varvarigos**

17-22 **Human capital threshold effects in economic development: A panel data approach with endogenous threshold, Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis**

18-22 Distributional aspects of rent seeking activities in a Real Business Cycle model, Tryfonas Christou, Apostolis Philippopoulos and Vanghelis Vassilatos

<u>2023</u>

01-23 Real interest rate and monetary policy in the post Bretton Woods United States, George C. Bitros and Mara Vidali

02-23 Debt targets and fiscal consolidation in a two-country HANK model: the case of Euro Area, Xiaoshan Chen, Spyridon Lazarakis and Petros Varthalitis

03-23 Central bank digital currencies: Foundational issues and prospects looking forward, George C. Bitros and Anastasios G. Malliaris

04-23 The State and the Economy of Modern Greece. Key Drivers from 1821 to the Present, George Alogoskoufis

05-23 Sparse spanning portfolios and under-diversification with second-order stochastic dominance, Stelios Arvanitis, Olivier Scaillet, Nikolas Topaloglou

06-23 What makes for survival? Key characteristics of Greek incubated early-stage startup(per)s during the Crisis: a multivariate and machine learning approach, Ioannis Besis, Ioanna Sapfo Pepelasis and Spiros Paraskevas

07-23 The Twin Deficits, Monetary Instability and Debt Crises in the History of Modern Greece, George Alogoskoufis

08-23 Dealing with endogenous regressors using copulas; on the problem of near multicollinearity, Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis

09-23 A machine learning approach to construct quarterly data on intangible investment for Eurozone, Angelos Alexopoulos and Petros Varthalitis

10-23 Asymmetries in Post-War Monetary Arrangements in Europe: From Bretton Woods to the Euro Area, George Alogoskoufis, Konstantinos Gravas and Laurent Jacque

11-23 Unanticipated Inflation, Unemployment Persistence and the New Keynesian Phillips Curve, George Alogoskoufis and Stelios Giannoulakis

12-23 Threshold Endogeneity in Threshold VARs: An Application to Monetary State Dependence, Dimitris Christopoulos, Peter McAdam and Elias Tzavalis

13-23 A DSGE Model for the European Unemployment Persistence, Konstantinos Giakas

14-23 Binary public decisions with a status quo: undominated mechanisms without coercion, Efthymios Athanasiou and Giacomo Valletta

15-23 Does Agents' learning explain deviations in the Euro Area between the Core and the Periphery? George Economides, Konstantinos Mavrigiannakis and Vanghelis Vassilatos

16-23 Mild Explocivity, Persistent Homology and Cryptocurrencies' Bubbles: An Empirical Exercise, Stelios Arvanitis and Michalis Detsis

17-23 A network and machine learning approach to detect Value Added Tax fraud, Angelos Alexopoulos, Petros Dellaportas, Stanley Gyoshev, Christos Kotsogiannis, Sofia C. Olhede, Trifon Pavkov

18-23 Time Varying Three Pass Regression Filter, Yiannis Dendramis, George Kapetanios, Massimiliano Marcellino

19-23 From debt arithmetic to fiscal sustainability and fiscal rules: Taking stock, George Economides, Natasha Miouli and Apostolis Philippopoulos

20-23 Stochastic Arbitrage Opportunities: Set Estimation and Statistical Testing, Stelios Arvanitis and Thierry Post

21-23 Behavioral Personae, Stochastic Dominance, and the Cryptocurrency Market, Stelios Arvanitis, Nikolas Topaloglou, and Georgios Tsomidis

22-23 Block Empirical Likelihood Inference for Stochastic Bounding: Large Deviations Asymptotics Under *m*-Dependence, Stelios Arvanitis and Nikolas Topaloglou

23-23 A Consolidation of the Neoclassical Macroeconomic Competitive General Equilibrium Theory via Keynesianism (Part 1 and Part 2), Angelos Angelopoulos

24-23 Limit Theory for Martingale Transforms with Heavy-Tailed Noise, Stelios Arvanitis and Alexandros Louka

2024

01-24 Market Timing & Predictive Complexity, Stelios Arvanitis, Foteini Kyriazi, Dimitrios Thomakos

02-24 Multi-Objective Frequentistic Model Averaging with an Application to Economic Growth, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, Nikolas Topaloglou

03-24 State dependent fiscal multipliers in a Small Open Economy, Xiaoshan Chen, Jilei Huang, Petros Varthalitis

04-24 Public debt consolidation: Aggregate and distributional implications in a small open economy of the Euro Area, Eleftherios-Theodoros Roumpanis

05-24 Intangible investment during the Global Financial Crisis in the EU, Vassiliki Dimakopoulou, Stelios Sakkas and Petros Varthalitis

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
SCHOOL OF
ECONOMIC
SCIENCES

ΤΜΗΜΑ
ΟΙΚΟΝΟΜΙΚΗΣ
ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF
ECONOMICS

Department of Economics
Athens University of Economics and Business

The Department is the oldest Department of Economics in Greece with a pioneering role in organising postgraduate studies in Economics since 1978. Its priority has always been to bring together highly qualified academics and top quality students. Faculty members specialize in a wide range of topics in economics, with teaching and research experience in world-class universities and publications in top academic journals.

The Department constantly strives to maintain its high level of research and teaching standards. It covers a wide range of economic studies in micro-and macroeconomic analysis, banking and finance, public and monetary economics, international and rural economics, labour economics, industrial organization and strategy, economics of the environment and natural resources, economic history and relevant quantitative tools of mathematics, statistics and econometrics.

Its undergraduate program attracts high quality students who, after successful completion of their studies, have excellent prospects for employment in the private and public sector, including areas such as business, banking, finance and advisory services. Also, graduates of the program have solid foundations in economics and related tools and are regularly admitted to top graduate programs internationally. Three specializations are offered:1. Economic Theory and Policy, 2. Business Economics and Finance and 3. International and European Economics. The postgraduate programs of the Department (M.Sc and Ph.D) are highly regarded and attract a large number of quality candidates every year.

For more information:

https://www.dept.aueb.gr/en/econ/