INDUSTRIAL ORGANIZATION, ORDER INTERNALIZATION, AND INVARIANCE

Albert S. Kyle
Anna A. Obizhaeva
Yajun Wang

Working Paper

No 274

NES Working Paper series

August

2020

Industrial Organization, Order Internalization, and Invariance

Albert S. Kyle, Anna A. Obizhaeva, and Yajun Wang

First Draft: May 24, 2014
This Draft: November 18, 2017*

Abstract

We present a one-period model of oligopolistic strategic trading among symmetric traders who agree to disagree about the precision of their private signals. We derive several invariance relationships relating the number of firms, number of firm's employees, average trade size, price impact, and pricing accuracy to dollar volume and returns volatility. Since a substantial part of order flow is often internalized within firms and does not reach the marketplace, invariance relationships can be modified to account for internalized order flow.

Keywords: invariance, agreement to disagree, market power, trade size, price impact, pricing accuracy, volume, volatility, market microstructure, industrial organization.

*Kyle: Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, akyle@rhsmith.umd.edu. Obizhaeva: New Economic School, 100A Novaya Street, Skolkovo, Moscow, Russia, 143026, aobizhaeva@nes.ru. Wang: Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, ywang22@rhsmith.umd.edu.

Even though market microstructure variables—such as volume, volatility, bet size, number of bets, liquidity measures, and pricing accuracy—vary significantly across assets and across time, they satisfy particular invariance relationships proposed by Kyle and Obizhaeva (2016). We derive these invariance relationships theoretically using a simple one-period equilibrium model of trading among symmetric oligopolistic firms. Our model also generates specific quantitative predictions about some industrial organization aspects of financial firms such as the number of financial firms, the number of traders employed at each firm, and assets under management.

Our model is based on a one-period model of Kyle, Obizhaeva and Wang (2017). There are informed firms and noise firms who trade a risky asset against a risk-free numeraire. Informed firms trade based on informative private signals. Noise firms trade based on uninformative signals, as if it were information; similar motivation for noise trading can be found in Treynor (1995) and Black (1986). Both informed and noise firms believe that they trades on informative signal, while some of the other firms trade on uninformative signals; however, trades of both types of firms are indistinguishable. Firms optimize their strategies taking into account its market power. In the symmetric equilibrium, prices are equal to the average of firms' risk-neutral buy-and-hold valuations, and each firm trades proportionally to the difference between its own valuation and the average valuation of other market participants that it infers from prices.

To generate invariance relationships, we make three additional adjustments to the baseline model. We model the endogenous choice by firms to acquire private information and impose two appropriate restrictions on the two easily observable variables such as volume and volatility.

First, firms decide to acquire private information based on their break-even condition so that profits of trading on information cover the cost of acquiring it. The equilibrium fraction of informed firms turns out to be close to a half with the approximation error being proportional to the degree of risk aversion and the cost of a private signal. If firms are risk averse and information is costly, then firms have fewer incentives to acquire information, and this effect pushes down the number of informed firms relative to the number of noise firms in the market.

Second, the distinctive property of invariance is the strong emphasis it puts on *trading volume*. Even though trading volume is one of the most important characteristics of financial markets, it has by and large slipped the attention of researchers, who instead usually tend to study asset prices; this property has been even reflected in the name of the field of asset pricing itself. Yet, volume is one of the few easily observable variables, and thus reasonable predictions concerning trading volume must be among main indicators of how successful a proposed model of financial markets is. Good models must explain why people trade so much. Their predictions must match empirically observed levels of trading volume. Our model based on the agreement of firms to disagree about their signals has a good concept of volume. We match theoretical volume to empirically observable volume, and the volume condition allows us later to endogenize either the number of firms or the number of traders employed by each firm.

Third, another easily observable and therefore important characteristic of the market is *price volatility*. In one-period models, several theoretical measures can be potentially used for volatility, since any one-period model has effectively two different price changes between pre-trade price and trade price as well as between trade price and post-trade price. This makes it difficult to use one-period models to derive realistic implications for continuously operating dynamic financial markets. Similar concerns are relevant for any non-stationary model. We believe it is more appropriate to think of volatility of the difference between pre-trade price and trade price as the proxy for price volatility, while volatility of the difference between trade price and post-trade liquidation value can be thought of as a proxy for pricing accuracy. This conceptual issue is especially important for derivation of invariance relationships, because market microstrucure invariance is ultimately a dynamic concept, which is closely related to how volume and volatility unfold over time. We match theoretical volatility to empirically observable volatility, and it essentially allows us to endogenize the volatility of fundamentals in the model.

With all aspects of the model carefully thought out, we derive invariance relationships as equilibrium properties of the model. These relationships also coincide with invariance relationships derived in Kyle and Obizhaeva (2016) as implications of empirical conjectures and derived in Kyle and Obizhaeva (2017*a*) based on the concepts of dimensional analysis and leverage neutrality. Kyle and Obizhaeva (2017*b*) shows how to derive invariance relationships in the context of a much more complicated dynamic equilibrium model, and we discuss how results of our model can be mapped into their results.

It is remarkable that even a simple stylized model is able to generate quantitative invariance predictions that match so well the empirical evidence. For example, Kyle and Obizhaeva (2016) document invariance relationships for the size distributions of portfolio transition orders. Kyle, Obizhaeva and Tuzun (2016) find similar relationship between the size distribution of transactions and trading activity in the Trade and Quote data set (TAQ). Andersen et al. (2016) report invariance relationships for number of transactions and average size of transactions in the data for S&P500 E-mini futures. Similarly, Kyle et al. (2014) study invariance relations for the number of news articles using Thomson-Reuters data. Bae et al. (2014) discuss an invariance relationship for the number of buy-sell switching points in the South Korean market.

The proof of invariance relationships relies on the approximation, which holds only if the number of firms is relatively large and the precision of private signals of each firm or their risk aversion are sufficiently small. This potentially identifies natural boundaries of invariance. For example, invariance relationships are less likely to be found in the market for financial securities in which only a few firms are active.

Our model also generates new results. Some order flow tends to be internalized in internal crossing systems of financial firms. Banks often internalize order flow of institutional investors. Brokers often internalize order flow of retail investors who trade using their systems. We show how to get invariance relationships by adjusting trading volume and trading activity for internalized order flow. This discussion highlights why it is empirically difficult to identify bets, or independent trading decisions, in the order flow. We show that the number of bets has to be calculated not only as the number of firms or trades, but rather as its product with the precision of private signals. The number of bets is therefore equal to the total number of units of precision of private information incorporated by the market. Another difficulty, discussed in Kyle and Obizhaeva (2016) and left for the future research, arises due to a common practice to shred bets—often correlated across several market participants—over time to minimize transaction costs.

Our model generates specific quantitative predictions about relationships between the number of financial firms, the number of traders employed at each firm (proxied by the precision of firm's private information), and assets under management (proxied by degree of risk aversion). Those results may be further developed to provide valuable theoretical guidance to the literature on industrial organization of financial sector, which studies the allocation of talent and capital across firms, career choice of finance professionals, and their wages. Murphy, Shleifer and Vishny (1991), Schwarzkopf and Farmer (2010), and Philippon and Reshef (2012) are examples of related research.

The remainder of the paper proceeds as follows. In Section 1 we present the model and solve the equilibrium with information acquisition and trading volume restriction. We derive invariance relationships for the two cases with endogenous number of firms and endogenous precision of signals produced by each firm in Section 2. We conclude in Section 3.

1 Trading Game

There are N financial firms that trade a risky asset with liquidation value $\tilde{v} \sim N(0, \tau_v^{-1})$ against a safe numeraire asset with a liquidation value of one. Each firm n where $n = 1, \ldots, N$ is endowed with a zero inventory of the risky asset. The risky asset is in zero net supply.

Each firm observes a public signal $\tilde{i}_0$ with precision τ_0 about a liquidation value,

$$\tilde{i}_0 := \tau_0^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v}) + \tilde{e}_0, \tag{1}$$

where $\tilde{e}_0 \sim N(0,1)$. Each firm n also observes a private signal $\tilde{i}_n$ with precision τ_n,

$$\tilde{i}_n := \tau_n^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v}) + \tilde{e}_n, \tag{2}$$

where $\tilde{e}_n \sim N(0,1)$. The cost of observing a public signal is zero, and the cost of generating a private signal i_n is equal to c_I dollars. Given the symmetry of the equilibrium, each firm n also infers from the market price the average of signals of the other firms

$$\tilde{i}_{-n} := \frac{1}{N-1} \sum_{m \neq n} \tilde{i}_m. \tag{3}$$

The asset payoff $\tilde{v}$, the public signal error $\tilde{e}_0$, and N private signal errors $\tilde{e}_1, \ldots, \tilde{e}_N$ are all independently distributed.

A particular choice of information structure is important for derivation of invariance relationships. We assume that each firm n spends some resources to generate a private signal. Firm's employers run multiple predictive regressions of past realizations of normalized variable of interest $\tau_v^{1/2} \cdot \tilde{v}$ on various combinations of factors they choose to test in constructing their strategy. Upon extensive analysis, they select a particular combination of factors $\tilde{i}_n$ that forecast the fundamental value the best. It can be shown that, when τ_n is small, it is approximately equal to the R-square of a regression in which the scaled variable $\tau_v^{1/2} \cdot \tilde{v}$ is regressed on the scaled signal $\tilde{i}_n$. Therefore the bigger is the precision τ_n of firm n's signal, the higher is the R-square of the predictive model and the bigger is its potential to generate profits. In other words, the precision τ_n can be viewed as a proxy for the number of traders employed by firm n who participate in generating firm n's signal.

When implementing the regression analysis, traders scale the dependent variable $\tilde{v}$ by its standard deviation of $\tau_v^{-1/2}$, which is assumed to be a common knowledge in the model. This scaling is consistent with a practice of standardizing variables in the financial industry. Indeed, variables are often measured in different units; for example, share volume and number of transactions may be used to forecast dollar bid-ask spread. Scaling ensures that precision of information becomes unitless and does not depend on the units in which variables are measured. The fact that τ_n does not depend on specific units in which $\tilde{v}$ is measured will be important for our results.

Firms agree about the precision of the public signal τ_0 and agree to disagree about the precisions of private signals τ_n. Each firm n thinks its private signal has a precision of $\tau_n = \tau_H$, and it pays c_I dollars to acquire it.

Each firm believes that $N-1$ other firms can be of the two types. There are $N_I - 1$ "informed" firms like firm n itself, and there are N_U "noise" firms. Informed firms generate private signals of precision τ_H at a cost c_I. Noise firms waste the same resources to generate private signals, but in reality their signals have no information, either because firm lack expertise or they are prone to making mistakes in their analysis; nevertheless noise firms trade on noise as it were information. Denote the fraction of informed firms in the market as

$$\theta := \frac{N_I - 1}{N_U + N_I - 1}. \tag{4}$$

No firm knows which of the firms are noise and which of them are informed; it is sure only about its own type.

Firms submit their demand schedules $\tilde{X}_n(p) := X_n(\tilde{i}_0, \tilde{i}_n, p)$ to an auctioneer of a single-price auction. An auctioneer then calculates a market clearing price $\tilde{p} := \tilde{p}[X_1, \ldots, X_N]$. Firm n's terminal wealth is

$$\tilde{w}_n := (\tilde{v} - \tilde{p}) \cdot \tilde{X}_n(\tilde{p}). \tag{5}$$

Each firm n chooses X_n to maximize the expected exponential utility of its terminal wealth

$$\max \mathrm{E}^n \left[-\mathrm{e}^{-A \cdot \tilde{w}_n} \right] \tag{6}$$

with a risk tolerance parameter $1/A$, using its own beliefs about precision of signals to calculate the expectation. The exponential utility function is a realistic assumption for modelling financial markets where asset owners—such as pension plans, endowments, and foundations—hire financial firms to manage their assets. A risk tolerance parameter $1/A$ can be thought of as a measure of how much capital asset owners delegate to firms to manage, i.e. how much risk firms are allowed to take at a discretion of an asset owner. Even when their portfolios gain or lose value, the capital delegated to firms may remain somewhat constant.

Firms are modelled as imperfect competitors. They explicitly take into account the effect of their trading on prices. In practice, the importance of imperfect competition is reflected in significant amount of resources devoted by the industry to developing good transaction costs models and designing cost efficient execution algorithms.

We define an equilibrium as a set of trading strategies $X_1, \ldots, X_N$ such that each trader's strategy maximizes his expected utility, taking as given the trading strategies of the other traders. This is equivalent to a Bayesian Nash equilibrium in trading strategies $X_1, \ldots, X_N$, except for the assumption that traders do not share a common prior. Traders are imperfect competitors who explicitly take into account that the price $\tilde{p}$ is a function of the quantity x_n they trade.

1.1 Conjectured Linear Strategies and Bayesian Updating

We seek to characterize equilibria with symmetric linear trading strategies. To do so, we use the "no-regret" approach, as in Kyle (1989), which assumes that each firm observes its residual linear supply schedule, infers the average of other firms' signals from the intercept of this schedule, picks the quantity x_n on this residual supply schedule which maximizes its expected utility. It can be shown then that the firm can implements this optimal choice x_n even in the more complicated original problem of choosing a demand schedule $x_n = X_n(i_0, i_n, p)$ based on prices, without first observing the residual supply schedule.

Firm n conjectures that the other $N-1$ firms submit symmetric linear demand schedules of the form

$$X_m(i_0, i_m, p) = \alpha \cdot i_0 + \beta \cdot i_m - \gamma \cdot p, \qquad m = 1, \ldots, N, \quad m \neq n. \tag{7}$$

Then, it infers from the market clearing condition

$$x_n + \sum_{m \neq n} (\alpha \cdot i_0 + \beta \cdot i_m - \gamma \cdot p) = 0 \tag{8}$$

that its residual supply schedule $P(\cdot)$ is a function of its quantity x_n given by

$$P(x_n) = \frac{\alpha}{\gamma} \cdot \tilde{i}_0 + \frac{\beta}{\gamma} \cdot \tilde{i}_{-n} + \frac{1}{(N-1)\gamma} \cdot x_n. \tag{9}$$

Since firm n observes the public signal $\tilde{i}_0$, its own inventory S_n, and the quantity it trades x_n, it can infer the average of other firms' signals $\tilde{i}_{-n}$ from observing the intercept $\alpha/\gamma \cdot \tilde{i}_0 + \beta/\gamma \cdot \tilde{i}_{-n}$ of its residual supply schedule.

Equation (9) yields the price impact of trading one share of a risky asset,

$$\lambda := \frac{1}{(N-1) \cdot \gamma}. \tag{10}$$

The price impact is small, when the number of firms N and the price sensitivity parameter γ are large, i.e., firms tend to provide a lot of liquidity to each other.

Let $\mathrm{E}_n[\ldots]$ and $\mathrm{Var}_n[\ldots]$ denote firm n's expectation and variance operators conditional on observing signals i_0, i_n, and i_{-n}. Using standard formulas for conditional means and variances of jointly normally distributed random variables, we calculate the total precision in the economy as

$$\tau := \left(\mathrm{Var}^n\left[\tilde{v}\right]\right)^{-1} = \tau_v \cdot \left(1 + \tau_0 + \tau_H + (N-1)\cdot\theta^2\cdot\tau_H\right), \tag{11}$$

and obtain

$$\mathrm{E}^n\left[\tilde{v}\right] = \frac{\tau_v^{1/2}}{\tau}\cdot\left(\tau_0^{1/2}\cdot\tilde{i_0} + \tau_H^{1/2}\cdot\tilde{i_n} + (N-1)\cdot\theta\cdot\tau_H^{1/2}\cdot\tilde{i}_{-n}\right). \tag{12}$$

Firm n assigns the precision τ_0 to a public signal, τ_H to its own signal and $\theta^2\tau_H$ to each of other $N-1$ signals. Each firm knows that, in addition to itself, there are $\theta\cdot(N-1)$ informed firms, but it does not know who gets informed signals and who gets noise signals; in formulas above, this reduces the effective precision of information revealed by private signals of the others from $\theta\cdot\tau_H$ to $\theta^2\cdot\tau_H$. If each firms knew precisely which of the other firms were informed and which of them were noise, then total precision would have $\theta\cdot\tau_H$.

Keeping the total number of firms N fixed, the smaller is the fraction of informed traders θ, the smaller is the weight assigned to signals of other traders. In the model with noise firms, each firm is overconfident about its own signal, and the overconfidence generates trading.

1.2 Utility Maximization with Market Power

Conditional on firm n's information, its terminal wealth $\tilde{w}_n$ is a normally distributed random variable with mean and variance given by

$$\mathrm{E}^n\left[\tilde{w}_n\right] = \mathrm{E}^n\left[\tilde{v}\right]\cdot x_n - P(x_n)\cdot x_n, \tag{13}$$

$$\mathrm{Var}^n\left[\tilde{w}_n\right] = x_n^2\cdot\mathrm{Var}^n\left[\tilde{v}\right]. \tag{14}$$

Normal distributions imply that expected utility is given by

$$\mathrm{E}^n\left[-\mathrm{e}^{-A\cdot\tilde{w}_n}\right] = -\exp\left(-A\cdot\mathrm{E}^n\left[\tilde{w}_n\right] + \tfrac{1}{2}A^2\cdot\mathrm{Var}^n\left[\tilde{w}_n\right]\right). \tag{15}$$

Maximizing this function is equivalent to maximizing the simpler function $\mathrm{E}^n\left[\tilde{w}_n\right] - \frac{1}{2}A\cdot\mathrm{Var}^n\left[\tilde{w}_n\right]$.

Oligopolistic firm n exercises its market power by taking into account how its quantity traded x_n affects the price $\tilde{P}(x_n)$ on its residual supply schedule (9). After plugging equations (11), (12), (13) and (14) into equation (15), firm n chooses the quantity to trade x_n that maximizes the following quadratic expression

$$\frac{\tau_v^{1/2}}{\tau} \cdot \left(\tau_0^{1/2} \cdot \tilde{i}_0 + \tau_H^{1/2} \cdot \tilde{i}_n + (N-1) \cdot \theta \cdot \tau_H^{1/2} \cdot \tilde{i}_{-n} \right) \cdot x_n - P(x_n) \cdot x_n - \frac{A}{2\tau} \cdot x_n^2. \tag{16}$$

Solution. Under the tentative assumption that firm n knows the value of i_{-n}, we plug equation (9) into equation (16) and use the first order condition to find its optimal demand,

$$x_n = \frac{\frac{\tau_v^{1/2}}{\tau} \cdot \left(\tau_0^{1/2} \cdot i_0 + \tau_H^{1/2} \cdot i_n + (N-1) \cdot \theta \cdot \tau_H^{1/2} \cdot i_{-n} \right) - \left(\frac{\alpha}{\gamma} \cdot i_0 + \frac{\beta}{\gamma} \cdot i_{-n} \right)}{\frac{2}{(N-1)\gamma} + \frac{A}{\tau}}. \tag{17}$$

In the numerator of this equation, the first term is firm n's expectation of the liquidation value and the second term is the market clearing price when firm n trades a quantity of zero. In the denominator, the first and second terms reflect how firm n restricts the quantity it trades due to its market power and risk aversion, respectively.

As in Kyle (1989), even though firm n does not observe i_{-n} explicitly, it is still able to implement this optimal point by a linear demand schedule, because the necessary sufficient statistics can be inferred from the market price. The strategy is ex post optimal. We next show how to implement this strategy by using limit orders, i.e., conditioning trade sizes on market price.

Define the constant

$$C := \frac{1}{(N-1) \cdot \gamma} + \frac{A}{\tau} + \frac{\theta \cdot \tau_H^{1/2} \cdot \tau_v^{1/2}}{\tau \cdot \beta}. \tag{18}$$

We solve for i_{-n} instead of p in the market clearing condition (8), substitute this solution into equation (17) above, and then solve for x_n to derive a demand schedule $X_n(i_0, i_n, p)$ for firm n as a function of price p,

$$\begin{aligned} X_n(i_0, i_n, p) = \frac{1}{C} \cdot \Bigg[& \frac{\tau_v^{1/2}}{\tau} \left(\tau_0^{1/2} - (N-1) \cdot \theta \cdot \tau_H^{1/2} \cdot \frac{\alpha}{\beta} \right) \cdot i_0 + \frac{\tau_H^{1/2}}{\tau} \cdot \tau_v^{1/2} \cdot i_n \\ & + \left(\frac{(N-1) \cdot \theta \cdot \tau_H^{1/2}}{\tau} \cdot \frac{\gamma}{\beta} \cdot \tau_v^{1/2} - 1 \right) \cdot p \Bigg]. \end{aligned} \tag{19}$$

In a symmetric linear equilibrium, the strategy chosen by firm n must be the same as the linear strategy (7) it conjectures for the other firms. Equating corresponding coefficients of variables i_0, i_n, and p yields the system of three equations in terms of three unknowns α, β, and γ:

$$\alpha = \frac{\tau_v^{1/2}}{C} \cdot \left(\frac{\tau_0^{1/2}}{\tau} - \frac{(N-1) \cdot \theta \cdot \tau_H^{1/2}}{\tau} \cdot \frac{\alpha}{\beta} \right), \qquad \beta = \frac{\tau_v^{1/2}}{C} \cdot \frac{\tau_H^{1/2}}{\tau}, \tag{20}$$

$$\gamma = -\frac{1}{C}\left(\frac{(N-1)\cdot\theta\cdot\tau_H^{1/2}}{\tau}\cdot\frac{\gamma}{\beta}\cdot\tau_v^{1/2}-1\right), \qquad \delta = \frac{1}{C}\left(\frac{\theta\cdot\tau_H^{1/2}}{\tau}\cdot\frac{\delta}{\beta}\cdot\tau_v^{1/2}+\frac{A}{\tau}\right). \tag{21}$$

The unique solution for α, β, and γ is

$$\beta = \frac{(N-2)-2(N-1)\cdot\theta}{A\cdot(N-1)}\cdot\tau_H^{1/2}\cdot\tau_v^{1/2}, \tag{22}$$

$$\alpha = \frac{\tau_0^{1/2}}{(1+(N-1)\theta)\cdot\tau_H^{1/2}}\cdot\beta, \qquad \gamma = \frac{\tau}{(1+(N-1)\theta)\cdot\tau_H^{1/2}}\cdot\frac{\beta}{\tau_v^{1/2}}, \qquad \delta = \frac{A}{(1-\theta)\cdot\tau_H^{1/2}}\cdot\frac{\beta}{\tau_v^{1/2}}. \tag{23}$$

Plugging these constants into equation (7), we derive the equilibrium strategy. The equilibrium price can be then derived from the market clearing condition.

Equilibrium Price and Quantities. There is always a trivial no-trade equilibrium. If each trader submits a no-trade demand schedule $X_n(.) \equiv 0$, then such a no-trade demand schedule is optimal for all traders. In this no-trade equilibrium, an auctioneer cannot establish a meaningful market price.

We next present the equilibrium with trade. Define the exogenous parameter θ^* by

$$\theta^* := \frac{1}{2} - \frac{1}{2(N-1)}. \tag{24}$$

The constant θ^* can be also written as $\theta^* = \frac{N-2}{2(N-1)}$. The fraction one-half reflects monopoly power over firm's own information. Monopolist restricts his trading when trading against a residual supply curve with a positive slope and prices never reveal more than a half of his precision. The fraction $(N-2)/(N-1)$ arises from the Cournot-like competition among $N-1$ traders competing to provide him liquidity. Both monopoly and "Cournot" effect get combined together in equilibrium calculation as a product. When N goes to infinity, θ^* converges to the monopolistic fraction 1/2. Similar constant can be found in Kyle (1989) and other industrial organization papers.

Define the exogenous measure of "disagreement" Δ by

$$\Delta := \frac{1}{\theta} - \frac{1}{\theta^*}. \tag{25}$$

When θ is small, then the measure of disagreement is large, because there are very few firms with informative signals.

Theorem 1. (The Equilibrium Trading Quantity and Price). *There exists a unique symmetric equilibrium with linear trading strategies and nonzero trade if and only if* $\Delta > 0$ *holds. The equilibrium satisfies the following:*

1. *The equilibrium trading strategy is linear in the disagreement between each firm and the rest of the market. The equilibrium quantity traded by trader n is*

$$x_n = \frac{2\tau_H^{1/2} \cdot (N-1)}{\tau_v^{-1/2} A \cdot N} \cdot \left(\theta^* - \theta\right) \cdot (\tilde{i}_n - \tilde{i}_{-n}); \tag{26}$$

2. *The equilibrium price is fully revealing and equal to the average of firms' risk-neutral buy-and-hold valuations,*

$$P = \frac{\tau_v^{1/2}}{\tau} \cdot \left(\tau_0^{1/2} \cdot \tilde{i}_0 + \frac{\tau_H^{1/2} + (N-1) \cdot \theta \tau_H^{1/2}}{N} \cdot \sum_{n=1}^{N} \tilde{i}_n\right). \tag{27}$$

The proof is in Appendix.

The symmetric equilibrium with the positive volume can be sustained only if $\Delta > 0$ or equivalently $\theta < \theta^*$, i.e., firms are sufficiently more than twice overconfident about the square root of the precision of their own signal comparing to the rest of the market. Indeed, each firm assigns the precision τ_H to its private signal, whereas the rest of the market effectively assigns to it the precision of $\theta^2 \tau_H$.

Note that the effective precision assigned to signals in the price equation (27) is equal to $\theta^2 \tau_H$ rather than "average" precision $\theta \tau_H$. Since $\theta < 1/2$, it implies that too little private information is incorporated into equilibrium price. This effect is somewhat similar to the paradox of the impossibility of informationally efficient markets of Grossman and Stiglitz (1980). It also complements the observation of Allen, Morris and Shin (2006) that the market tends to overreact to public information and underreact to private information.

Firms' Decision to Acquire Private Information. Next, we consider the decision of firm n to spend an exogenously specified cost c_I on generating a private signal and derive the "entry condition."

Suppose that firm n starts with a capital of $w_{0,n}$ in cash. Before trading begins, it has to decide on whether to buy a private signal and participate in trading. A default option is to do nothing and just keep cash in a bank account. The second option is to pay out c_I dollars for a private signal of precision τ_H, trade on it, and make $w_{0,n} - c_I + (\tilde{v} - p) \cdot \tilde{x}_n$.

Let $\mathrm{E}_0^n[\dots]$ denote unconditional expectation of firm n prior to observing a signal. In the equilibrium, the two options have to be equivalent in utility terms,

$$-\mathrm{e}^{-A\cdot(w_{0,n})} = \mathrm{E}_0^n\left[-\mathrm{e}^{-A\cdot(w_{0,n}-c_I+(\tilde{v}-p)\cdot\tilde{x}_n)}\right]. \tag{28}$$

This equation further yields,

$$\mathrm{e}^{-A\cdot c_I} = \mathrm{E}_0^n\left[\mathrm{e}^{-A\cdot\mathrm{E}^n[\tilde{v}-p]\cdot\tilde{x}_n+A^2\cdot\tilde{x}_n^2\cdot\mathrm{Var}^n[\tilde{v}-p]/2}\right]. \tag{29}$$

Given the equilibrium price in equation (27), it can be shown that

$$\mathrm{E}^n[\tilde{v}-p] = (1-\theta)\cdot\tau_H^{1/2}\cdot\frac{\tau_v^{1/2}}{\tau}\cdot\frac{N-1}{N}\cdot(i_n-i_{-n}), \tag{30}$$

$$\mathrm{Var}^n[\tilde{v}-p] = \tau^{-1}. \tag{31}$$

Substituting the optimal trading strategy (26) and the two equations above into the break-even condition (29) yields

$$\mathrm{e}^{-A\cdot c_I} = \mathrm{E}_0^n\left[\mathrm{e}^{-\frac{\tau_v\tau_H}{\tau}\cdot\frac{(N-1)}{N}\cdot(\theta^*-\theta)\cdot(i_n-i_{-n})^2}\right]. \tag{32}$$

Since $i_n - i_{-n}$ is a normal random variable with a zero mean, we can apply the formula for the moment-generating function of χ^2 random variable.[1] This further yields

$$\mathrm{e}^{-A\cdot c_I} = \left(1+\frac{2\tau_v\tau_H}{\tau}\cdot\frac{(N-1)}{N}\cdot(\theta^*-\theta)\cdot\mathrm{Var}_0^n[i_n-i_{-n}]\right)^{-1/2}. \tag{33}$$

In the equation above, the total precision τ is defined in (11), and the variance of a normal random variable $i_n - i_{-n}$ is given by

$$\mathrm{Var}_0^n[i_n-i_{-n}] = 1+\tau_H\cdot(1-\theta)^2+(N-1)^{-1}. \tag{34}$$

$$\mathrm{Var}_0^n[i_n-i_{-n}] = \frac{N}{N-1}+\tau_H\cdot(1-\theta)^2. \tag{35}$$

[1] If $\tilde{u}$ is a normal random variable with a zero mean and a variance σ_u^2, then $\mathrm{E}\mathrm{e}^{a\cdot u^2} = (1-2\cdot a\cdot\sigma_u^2)^{-1/2}$ for any constant $a \le \frac{1}{2\sigma_u^2}$.

We can therefore further simplify the break-even condition:

$$c_I = \frac{1}{2\cdot A}\cdot \ln\Big(1+2\tau_H\cdot\frac{(N-1)}{N}\cdot\frac{1+\tau_H\cdot(1-\theta)^2+(N-1)^{-1}}{1+\tau_0+\tau_H+(N-1)\cdot\theta^2\cdot\tau_H}\cdot(\theta^*-\theta)\Big). \tag{36}$$

Since τ_H is very small, i.e., a small amount of information arrives in each time period like in continuous-time approximation, the first order approximation for *the entry condition* is

$$A\cdot c_I \approx \tau_H\cdot\frac{(N-1)}{N}\cdot\frac{1+\tau_H\cdot(1-\theta)^2+(N-1)^{-1}}{1+\tau_0+\tau_H+(N-1)\cdot\theta^2\cdot\tau_H}\cdot(\theta^*-\theta). \tag{37}$$

This equation yields solution for θ, which determines the ratio of the number of informed firms N_I and the number of noise firms N_U in the market. However, this is a complicated non-linear equation in terms of θ.

As the firms move closer to risk neutrality (i.e., risk tolerance $1/A \to \infty$), firms believe in increasingly better profit opportunities in the market, and more of them choose to spend resources on acquiring private signals, thus reducing the amount of relative overconfidence in the market $1/\theta$. In the limit, the parameter θ converges to a critical value of θ^* (with large N, $\theta\to\theta^*=\frac{1}{2}$) and the measure of disagreement $\Delta = 1/\theta - 1/\theta^*$ converges to zero; the equilibrium converges to a no-trade equilibrium.

Volume Equation. One of the key concepts in financial markets is the concept of trading volume. Since it is essential for the derivation of invariance relationships, we next derive the volume equation for our model. It is straightforward to show an auxiliary result.

Lemma 1. *Given beliefs of firm n, the variance of a normal random variable $\tilde{i}_m - \tilde{i}_{-m}$ is equal to*

$$\mathrm{Var}_0^n\left[\tilde{i}_m-\tilde{i}_{-m}\right] = \begin{cases} 1+(N-1)^{-1}+\tau_H\cdot(1-\theta)^2, & \textit{if } m=n,\\ 1+(N-1)^{-1}+\tau_H\cdot(1-\theta)^2\cdot(N-1)^{-2}, & \textit{if } m\neq n.\end{cases} \tag{38}$$

Let V denote an expected trading volume in shares. Due to symmetry, all firms will agree on expected trading volume.

$$\sum_{m=1}^{N}\mathrm{E}_0^n\left[|x_m|\right] = 2\cdot V. \tag{39}$$

Since each transaction corresponds to a buy order matching a sell order, a scalar 2 ensures that we double count observable trading volume to uncover bet volume.[2]

[2] Kyle and Obizhaeva (2016) distinguish the concepts of bet volume $\bar{V}$ and trading volume V. The relation be-

Taking into account the optimal trading strategies (26), we get ***the volume equation***:

$$\sqrt{2/\pi}\cdot\frac{2\tau_H^{1/2}\cdot(N-1)}{\tau_v^{-1/2}A\cdot N}\cdot(\theta^*-\theta)\cdot\sum_{m=1}^{N}\Big(\mathrm{Var}_0^n\left[i_m-i_{-m}\right]\Big)^{1/2}=2\cdot V. \tag{40}$$

A scalar $\sqrt{2/\pi}$ converts the expectation of the absolute value of a normal random variables into its standard deviation.[3] Equation (38) further allows us to express the multiplier $\sum_{m=1}^{N}(\mathrm{Var}_0^n\left[i_m-i_{-m}\right])^{1/2}$ in terms of θ, τ_H, and N. Given τ_H and θ, we can use the volume equation to infer the number of firms N.

Trading Activity Equation. One of the important variables in market microstructure invariance is the trading activity of the risky security. As in Kyle and Obizhaeva (2016), the trading activity, denoted W, is defined as the product of share volume and dollar price volatility.

$$W := 2\cdot V\cdot\sigma_p, \tag{41}$$

where σ_p denote the price volatility.[4] Regardless of their beliefs, all firms will agree on expected trading activity. Intuitively, financial markets are about transferring risks. The measure of trading activity captures the aggregate amount of dollar risk transferred per game among traders. The bigger are trading volume and volatility, the more active is the market. For example, the markets of E-mini S&P500 index future, oil futures, or currency futures on USD and Euro pair have large volume and large volatility, being some of the most active markets in the world. The market of the U.S. Treasuries also have very large volume but much less risk transferred per dollar traded, so this market has less of trading activity. Small stocks with small trading volume have very little trading equation activity, even despite their somewhat high volatility relative to volatility of index futures.

To calculate trading activity, we need to know volume, discussed already above, and dollar volatility. The important conceptual question is what is a good proxy for volatility of price changes in the context of our one-period model. This proxy has to be consistent with volatility that one would get in a similar dynamic model. Similar issue arises whenever one-period

tween these variable is $\bar{V}=2/\zeta V$, where ζ is a volume or intermediation multiplier. When there are no intermediaries, as in our model, then the volume multiplier $\zeta=1$, i.e., bet volume is twice larger than trading volume, $\bar{V}=2V$, since each transaction consists of two bets, a buy bet and a sell bet. In Kyle and Obizhaeva (2017*b*), where each bet of traders is intermediated by a market maker, the volume multiplier is equal to two, i.e., bet volume is equal to trading volume, $\bar{V}=V$.

[3] If $\tilde{u}$ is a normal random variable with a zero mean and a variance σ_u^2, then $\mathrm{E}\left[|\tilde{u}|\right]=\sqrt{2/\pi}\sigma_u$.

[4] Kyle and Obizhaeva (2016), the trading activity is similarly defined as the product of dollar volume $V\cdot P$ and returns volatility σ.

models are used to derive prediction about dynamic markets.

In the one-period model, there are effectively two periods and three different price changes: price changes between pre-trade price $p_0 = 0$ and trade price $\tilde{p}$, price changes between trade price $\tilde{p}$ and the post-trade fundamental value $\tilde{v}$, and price changes between pre-trade prices $p_0 = 0$ and post trade fundamental values $\tilde{v}$. All three price changes have very different properties and have very different standard deviations. The predictions are quite different for two periods. For example, the first-period volatility may be lower or higher than the second-period volatility depending on parameters. This makes it difficult to use one-period models to derive realistic implications for continuously operating financial markets. Similar concerns are relevant for any non-stationary model.

Intuitively, the correct measure of volatility in a one-period model corresponds to the volatility of price change between pre-trade price and trade price:

$$\sigma_p := \left(\mathrm{Var}_0^n\left[p\right]\right)^{1/2}. \tag{42}$$

Moreover, let Σ denote the variance of the difference between trade price and fundamental value,

$$\Sigma := \mathrm{Var}_0^n\left[\tilde{v} - p\right] = \tau^{-1}. \tag{43}$$

It corresponds to the expected profits on buying a risky asset at price p and liquidating position in the far distant future at an unknown random value $\tilde{v}$. Then, $\Sigma^{-1/2}$ may be thought as a good measure of the pricing accuracy, as defined in Kyle and Obizhaeva (2017*b*).

Using equation (27), it can be shown that price volatility σ_p is proportional to fundamental volatility $\tau_v^{-1/2}$,

$$\sigma_p = \tau_v^{-1/2} \cdot \psi, \tag{44}$$

where a proportionality constant ψ is defined as

$$\psi = \frac{(\tau_0 + \tau_H(1 + (N-1)\cdot\theta)^2/N)^{1/2} \cdot (1 + \tau_0 + \tau_H(1 + (N-1)\cdot\theta)^2/N)^{1/2}}{1 + \tau_0 + \tau_H + (N-1)\cdot\theta^2\cdot\tau_H}. \tag{45}$$

Since we use a one-period model to generate intuition about dynamic markets and there is only a little of information revealed each period, it is reasonable to assume that τ_0, τ_H, and $N\cdot\theta^2\tau_H$ are small relative to one. Then pricing accuracy Σ and variance σ_p^2 satisfy the approximations

$$\Sigma = \tau^{-1} \approx \tau_v^{-1}, \qquad \sigma_p^2 \approx \tau_V^{-1}\cdot\tau_H\cdot(1 + (N-1)\cdot\theta)^2/N. \tag{46}$$

This is consistent with the intuition. When number of traders N, precision of information τ_H, and variance of fundamentals τ_v^{-1} are large, then a lot of information is incorporated into price through trading, leading to high price volatility. In dynamic steady-state models, σ_p^2 is essentially proportional to Σ, and the question about the volatility of which price change to use does not arise.

Note that one "problem" with this approach is the variance which goes into firms' demand functions. If they liquidate their position by collecting dividends, then the variance will be a number different from $1/\tau$. Another problem is the fraction of their information which they will capture when they liquidate.

Then, plug the volume equation (40) and the dollar volatility equation (44) into the equation (41) for trading activity to get:

$$\sqrt{2/\pi}\cdot\frac{2\tau_H^{1/2}\cdot(N-1)}{A\cdot N}\cdot\left(\theta^*-\theta\right)\cdot\sum_{m=1}^{N}\left(\mathrm{Var}_0^n\left[\tilde{i}_m-\tilde{i}_{-m}\right]\right)^{1/2}\cdot\psi=W, \tag{47}$$

where $\sum_{m=1}^{N}(\mathrm{Var}_0^n[\tilde{i}_m-\tilde{i}_{-m}])$ is defined in (38) and ψ is defined in (45); both are non-linear functions of θ, N, and τ_H. The transformation of volume into trading activity allows us to get rid of the unobservable parameter τ_v. We essentially relate unobservable fundamental volatility to easily observable price volatility.

2 Endogenous Invariance Relationships

Market microstructure invariance is based on the idea of inferring difficult to observe variables such as the price impact and the structure of order flow from more easily observable quantities such as volatility and volume. We next derive invariance results for the two cases with endogenous number of firms and endogenous information precision that each firm produces. To derive these invariance results, we need to impose the following assumption.

Assumption 1. *N is sufficiently large, τ_0, τ_H, and $N\cdot\theta^2\tau_H$ are small.*

2.1 Case with Endogenous Number of Firms

Suppose the precision of a private signal τ_H is fixed, but there is an endogenous number of firms N. This case essentially corresponds to the situation when there can be different number of firms trading in different risky assets, but the number of traders employed at each firms is constant. For the derivation of invariance results, the two key equations are the entry condition (37) and the equation for trading activity (47). These two equations make up the system of two non-linear equations in two unknown parameters N and θ, where A and c_I are exogenously given constants.

The system can be simplified under Assumption 1. If $\tau_0 \approx 0$, then most of information comes from trading. The entry condition (37) implies that $\frac{(N-1)}{AN} \cdot (\theta^* - \theta) \cdot \tau_H$ is equal to c_I. The equation (47) for trading activity then implies that $\sqrt{2/\pi} \cdot c_I \cdot 2\theta \cdot N^{3/2}$ is approximately equal to W. This further yields a solution for N and θ:

$$\theta \approx \frac{1}{2} - \frac{A \cdot c_I}{\tau_H}. \tag{48}$$

$$N \approx \left(\frac{W}{\sqrt{2/\pi} \cdot c_I \cdot 2\theta} \right)^{2/3}. \tag{49}$$

The number of traders N scales with 2/3 power of trading activity W. Using these results, it is easy to show that several other invariance relationships must hold.

Theorem 1 (**Invariance with Endogenous** N)**.** *If the number of firms N is endogenous, then under Assumption 1,*

$$N = \left(\frac{\lambda \cdot 2V}{\sigma_p \cdot \sqrt{2/\pi}} \right)^2 = \left(\frac{\mathrm{E}_0\left[|\tilde{x}|\right]}{2V} \right)^{-1} = \frac{\sigma_p^2}{\theta^2 \cdot \Sigma \cdot \tau_H} = \left(\frac{W}{\sqrt{2/\pi} \cdot c_I \cdot 2\theta} \right)^{2/3}, \tag{50}$$

and the distribution of microstructure invariant $\tilde{I}$ is given by

$$\tilde{I} := \frac{\tilde{x} \cdot (\mathrm{Var}_0\left[p\right])^{1/2}}{N^{1/2}} \approx c_I \cdot 2\theta \cdot (\tilde{i}_n - \tilde{i}_{-n}), \tag{51}$$

where $|\tilde{x}| := \frac{1}{N} \sum_{m=1}^{N} \mathrm{E}_0^n\left[|x_m|\right]$ is the average size of bets.

Proof. First, using equations (39) and (49), we find that the average size of bets $\mathrm{E}_0\left[|\tilde{x}|\right] := 2V/N$ in a symmetric model, as a fraction of bet volume $2V$, must satisfy:

$$\frac{\mathrm{E}_0\left[|\tilde{x}|\right]}{2 \cdot V} \approx \left(\frac{W}{\sqrt{2/\pi} \cdot c_I \cdot 2\theta} \right)^{-2/3}. \tag{52}$$

Second, under Assumption 1, equations (44) and (47) imply that

$$\sigma_p \approx \frac{\tau_H^{1/2}}{\tau_v^{1/2}} \cdot \frac{1 + (N-1)\theta}{N^{1/2}}, \qquad 2V \approx \frac{\sqrt{2/\pi} \cdot \tau_H^{1/2} \cdot (N-1) \cdot \tau_v^{1/2} \cdot (1 - 2\theta)}{A}. \tag{53}$$

From equations (10) and (53), it follows that

$$\frac{\lambda \cdot 2V}{\sigma_p \cdot \sqrt{2/\pi}} \approx N^{1/2}. \tag{54}$$

Using equation (49), we find that the price impact per one percent of bet volume $2V$ in volatility units $(\mathrm{Var}_0[p])^{1/2}$ must satisfy:

$$\frac{\lambda \cdot 2V}{\sigma_p \cdot \sqrt{2/\pi}} \approx \left(\frac{W}{\sqrt{2/\pi} \cdot c_I \cdot 2\theta} \right)^{1/3}. \tag{55}$$

Third, using equation (44) under Assumption 1, we have

$$\frac{\sigma_p^2}{\theta^2 \cdot \Sigma} \approx \frac{\tau_H \cdot (1 + (N-1)\theta)^2}{\theta^2 \cdot N} \approx \tau_H \cdot N. \tag{56}$$

Using equations (56) and (49), we find that the pricing accuracy Σ in units of volatility $\mathrm{Var}^{1/2}[p]$ must satisfy:

$$\frac{\Sigma}{\sigma_p^2} \approx \frac{1}{\theta^2 \tau_H} \cdot \left(\frac{W}{\sqrt{2/\pi} \cdot c_I \cdot 2\theta} \right)^{-2/3}. \tag{57}$$

Combining equation (49), (52), (55), and (57), we get equation (50).

Fourth, the market microstructure invariant $\tilde{I}$ is a random variable introduced by Kyle and Obizhaeva (2016). It measures the dollar risk transferred by a trade $\tilde{x}$ and measured in *business* time, which is adapted to arrival of bets, $(\mathrm{Var}_0[p]/N)^{1/2}$. Using the equations for equilibrium trade size (26), price volatility (44), as well as equations (48) and (49), we find that this random variable must have the following invariant distribution

$$\tilde{I} := \frac{\tilde{x} \cdot (\mathrm{Var}_0[p])^{1/2}}{N^{1/2}} \approx \tau_H \cdot (N-1) \cdot (1-2\theta) \cdot (1 + (N-1)\theta) \cdot (\tilde{i}_n - \tilde{i}_{-n}) \approx c_I \cdot 2\theta \cdot (\tilde{i}_n - \tilde{i}_{-n}). \tag{58}$$

□

These equations (50) and (51) in our one-period model are very similar to the invariance relationships presented in "Invariance Theorem" of the structural dynamic model of Kyle and Obizhaeva (2017*b*). That model has a more complicated dynamic structure, but it shares many similar modelling assumptions. The fundamental value evolves according to the geometric Brownian motion. Informed traders and noise traders arrive in the market randomly and trade with competitive risk-neutral market markers, who determine a break-even price. Noise traders turn over an exogenously specified fraction of a float by trading on uninformative signals. As in our model, informed traders choose to acquire informative signals if they think that it will be possible to recoup fixed and exogenously specified costs they spent on acquiring signals. Informed and noise traders can trade only once; even though signals differ in their information content, their distribution is approximately indistinguishable from each other.

Since all market participants in Kyle and Obizhaeva (2017*b*) are risk-neutral, the equilibrium ratio of informed and noise traders θ in that model is equal to $1/2$. In contrast in our model, market participants are risk averse. Equation (48) shows that risk aversion leads to a bigger degree of disagreement $\Delta = 1/\theta - 1/\theta^*$ in the market in the equilibrium. When firms become more risk averse, they value risky profit opportunities less, fewer of them find it optimal to acquire private signals ex ante, and the disagreement increases. In the other limiting case when firms are risk neutral, i.e., $A \to 0$, we get $\theta \to 1/2$ (keeping N large).

When N is large and $A \to 0$, both models generate the same invariance relationships. The equivalence can be established by noticing the natural correspondence between the parameters $\mathrm{Var}[\tilde{i}_n - \tilde{i}_{-n}]$ and c_I in our model to the parameters $\mathrm{Var}[\tilde{i}]$, moment ratio m, and expected dollar cost of a bet C_B in the structural model of Kyle and Obizhaeva (2017*b*). The variance of information is the same in both models, i.e., $\mathrm{Var}[\tilde{i}] = \mathrm{Var}[\tilde{i}_n - \tilde{i}_{-n}] = 1$. The cost of generating a private signal is the same, $C_B = c_I$, and the moment ratio that converts standard deviation into expectation of absolute value is the same, $m = \sqrt{2/\pi}$. Theorem is equivalent to Invariance Theorem 2 in Kyle and Obizhaeva (2017*b*).

The next theorem summarize all invariance predictions.

Theorem 2 (**Limit Case with Endogenous** N)**.** *When N is large and $A \to 0$, we get $\theta \to 1/2$ and*

$$N \sim W^{2/3}, \qquad \lambda \sim \frac{\sigma_p}{V} \cdot W^{1/3}, \qquad \frac{\mathrm{E}[|x|]}{V} \sim W^{-2/3}, \qquad \Sigma^{1/2} \sim \sigma_p \cdot W^{-1/3}. \tag{59}$$

We can interpret N as the number of bets and $\mathrm{E}[|x|]$ as their size. Then, these equations show how hard-to-observe microscopic parameters of the model—number of bets N, bet size $\mathrm{E}[|x|]$, price impact λ, and pricing accuracy $\Sigma^{1/2}$ scale with easily observable macroscopic parameter of trading activity W.

2.2 Case with Endogenous Precision of Signals

Suppose next that the number of firms N is fixed, but each firm can endogenously choose precision of its private signal τ_H. In other words, firms can choose the number of traders to hire or "units" of talent to employ. This case essentially corresponds to the situation when there is the same number of firms trading in different risky assets, but the number of resources that firms decide to allocate to each market differs across assets. To generate invariance relationships with respect to endogenous τ_H, several adjustments are necessary.

First, it is natural to assume that the cost of generating a signal i_n is proportional to its precision and equal to $\bar{c}_I \cdot \tau_H$, where $\bar{c}_I$ is a constant dollar cost per unit of precision. The firm n can hire several traders and compensate them proportionally to their marginal contribution to the design of an overall trading strategy. For example, the cost $\bar{c}_I \cdot \tau_H$ of generating a signal of precision τ_H can be interpreted as the cost of hiring M equally skilled traders, each of whom is paid $\bar{c}_I \cdot \tau_H/M$ dollars to generate a new piece of information of precision τ_H/M. Or, the same cost $\bar{c}_I \cdot \tau_H$ can be interpreted as the cost of hiring one very skilled trader generating a signal of precision $\tau_H/2$ for $\bar{c}_I \cdot \tau_H/2$ dollars and $M-1$ less skilled traders generating $M-1$ signals of precision $\tau_H/(2(M-1))$ for $\bar{c}_I \cdot \tau_n/(2(M-1))$ dollars each. In the language of "Fundamental Law of Active Management" of Grinold (1989), each trader seeks to generate an independent factor and increase the number of independent bets made by the firm.

Second, we need to think more carefully about the relevant concept of trading volume. Since each firm now consists of several traders who generate different investment ideas and often suggest to trade in opposite directions, a lot of trading will be internalized within each firm. If there are τ_H traders in a firm, then a market trade $\tilde{x}$ by each firm is the result of internalization of within-firm bets $\tilde{y}$ of τ_H employees of this firm,

$$\tilde{x} = \sum_{m=1}^{\tau_H} \tilde{y}_m. \tag{60}$$

Since $\tilde{x}$ and $\tilde{y}_m$ are random variable, we get

$$\mathrm{E}\left[|\tilde{x}|\right] = \tau_H^{1/2} \cdot \mathrm{E}\left[|y_m|\right]. \tag{61}$$

Thus, even though we observe N trade of size $\tilde{x}$, there are in reality $N \cdot \tau_H$ bets $\tilde{y}$ of size

$$\mathrm{E}\left[|\tilde{y}|\right] = \tau_H^{-1/2} \cdot \mathrm{E}\left[|\tilde{x}|\right]. \tag{62}$$

The true concept of volume represents not only open market volume, but also volume internalized within each firm. Let $\hat{V}$ denote this aggregate volume. Since internal volume within each firm is $\tau_H \cdot \mathrm{E}\left[|y|\right]$, we get

$$\hat{V} = N \cdot \tau_H \cdot \mathrm{E}\left[|y|\right] = N \cdot \tau_H^{1/2} \cdot \mathrm{E}\left[|\tilde{x}|\right] = \tau_H^{1/2} \cdot V. \tag{63}$$

One unit of expected volume V executed at the open market will correspond to $\tau_H^{1/2}$ units of expected total volume. The internalization multiplier is equal to the square root of the number of traders because some of the trades will cancel with each other in the internal pool of liquidity .

For example, suppose there are N firms, open market volume is V, and trade size is x. If each firm employs 100 of identically skilled traders, then size of their bets is $x/10$, number of their bets is $N \cdot 100$, total volume is $V \cdot 10$, if one accounts for order flow internalization within each firm.

The relevant measure of trading activity will need to be adjusted accordingly as well. Denote the total trading activity $\hat{W}$ being the product of daily volatility and total trading volume. Then,

$$\hat{V} = V \cdot \tau_H^{1/2} \quad \text{and} \quad \hat{W} = W \cdot \tau_H^{1/2}. \tag{64}$$

With these two adjustments, the two equations—the entry condition (37) and the equation for trading activity (47)—make up the system of two non-linear equations in two unknown parameters τ_H and θ, where A and $\bar{c}_I$ are exogenously given constants. Solving for θ and $\tau_H \cdot N$ yields:

$$\theta \approx \frac{1}{2} - A \cdot \bar{c}_I, \tag{65}$$

$$\tau_H \cdot N \approx \left(\sqrt{2/\pi} \cdot \bar{c}_I \cdot 2\theta\right)^{-2/3} \cdot \hat{W}^{2/3}. \tag{66}$$

The relative disagreement $1/\theta$ increases in aggregate market's risk aversion A/N and the total cost per unit of precision $\bar{c}_I \cdot N$. The total amount of information precision $\tau_H \cdot N$ scales with $2/3$ power of trading activity W. Other invariance relationships can be summarized as follows.

Theorem 3 (**Invariance with Endogenous** τ_H)**.** *If the precision of signal τ_H produced by firms is endogenous, then*

$$\tau_H \cdot N = \left(\frac{\lambda \cdot 2\hat{V}}{\sigma_p \cdot \sqrt{2/\pi}}\right)^2 = \left(\frac{\mathrm{E}\left[|\tau_H^{-1/2} \cdot \tilde{x}|\right]}{2\hat{V}}\right)^{-1} = \frac{\sigma_p^2}{\theta^2 \cdot \Sigma} = \left(\frac{\hat{W}}{\sqrt{2/\pi} \cdot \bar{c}_I \cdot 2\theta}\right)^{2/3}, \tag{67}$$

and the risk transferred by a trade $\tilde{x}$ in business time $\tilde{I}$ satisfies the following equation:

$$\tilde{I} := \frac{(\tau_H^{-1/2} \cdot \tilde{x}) \cdot \sigma_p}{(\tau_H \cdot N)^{1/2}} \approx \bar{c}_I \cdot 2\theta \cdot (\tilde{i}_n - \tilde{i}_{-n}). \tag{68}$$

These formulas show how to think about bets when each firm employs several traders generating investment ideas and some of their trading is internalized. It is natural to think about the number of bets as $\tau_H \cdot N$ and effective bet size as $\tau_H^{-1/2} \cdot \tilde{x}$. The volume and trading activity need to be adjusted for internalization of order flow as $\hat{V} = \tau_H^{1/2} \cdot V$ and $\hat{W} = \tau_H^{1/2} \cdot W$, respectively. A more careful thinking about internalization of bets is especially important nowadays when financial firms tend to have multiple trading desks and implement complicated internal systems of order flow internalization.

These predictions can be tested by studying the data on the number of financial firms trading in particular risky assets (proxy for N), the number of professional traders those firms hire (proxy for τ_H), and the size of assets under their management (proxy for A).

Our model also allows us to analyze implications of different industrial organization structures of financial industry. It is interesting to think about how financial markets will change if firms choose to merge with each others or, alternatively, split into several entities. As a specific example, suppose that there are N firms with τ_H professional employees. If each of the two firms merge, then there will be $N/2$ firms with $\tau_H \cdot 2$ professional employees each. The price volatility $\sigma_p^2 \approx \tau_v^{-1} \cdot \tau_H \cdot (1 + (N-1) \cdot \theta)^2 / N$ will remain approximately constant. The trading activity $W \approx \sqrt{2/\pi} \cdot \bar{c}_I \cdot \tau_H \cdot 2\theta \cdot N^{3/2}$ will decrease by a factor of $\sqrt{2}$, since more trading volume will be internalized within bigger firms. However, the trading activity adjusted for order flow internalization $\hat{W} = W \cdot \tau_H^{1/2}$, the total number of bets $\tau_H \cdot N$, and the average bet size $\tau_H^{-1/2} \cdot \tilde{x}$ will remain the same.

The discussion above does not take into account that merges will lead not only to changes in firms' precision of private information but also to changes in the amount of assets under their management. For example, merges between two firms will result in bigger firms managing double the amount of assets. In the framework with exponential utility functions, the merge of two firms with risk aversion A will result in a firm with risk aversion $A/2$, i.e., firms will effectively become less risk averse. From (65), θ will increase, further implying a lower degree of disagreement in the market. The change in risk aversion will have some effect on the aforementioned implications due to changes in θ, but those adjustments will be insignificant, when A is sufficiently small and θ is very close to $1/2$.

Theorem 4 (**Limit Case with Endogenous** τ_H)**.** *When N is large and $A \to 0$, we get $\theta \to 1/2$ and*

$$\tau_H \cdot N \sim \hat{W}^{2/3}, \qquad \lambda \sim \frac{\sigma_p}{\hat{V}} \cdot \hat{W}^{1/3}, \qquad \frac{\mathrm{E}\left[|\tau_H^{-1/2} \cdot \tilde{x}|\right]}{\hat{V}} \sim \hat{W}^{-2/3}, \qquad \Sigma^{1/2} \sim \sigma_p \cdot \hat{W}^{-1/3}. \tag{69}$$

These are invariance relationships, if one takes into account internalization of bet flow within firms with τ_H employees.

3 Conclusions

This paper provides an illustration of how invariance relationships can be derived in the context of a very simple one-period model. It is remarkable nevertheless how much quantitative predictions of this stylized model are consistent with existing empirical evidence. We think of this model as the simple illustration of how invariance relationships arise in the context of equilibrium models.

A good model of financial markets, however, has to be inherently dynamic. It must rely on sufficiently tractable, but yet realistic modelling of essential features of financial markets. These features include dynamically evolving public and private information as well as strategic trading by market participants who optimally manage transaction costs by shredding their orders over time. A fully-fledged dynamic model would generate a much richer set of implications concerning the dynamics of prices, order flow, volume, returns volatility, and inventories, yet we expect those predictions remain broadly consistent with market microstructure invariance principles of Kyle and Obizhaeva (2016). We leave the task of developing such a model for the future research.

References

Allen, Franklin, Stephen Morris, and Hyun Song Shin. 2006. "Beauty Contests and Iterated Expectations in Asset Markets." *Review of Financial Studies*, 19(3): 719–752.

Andersen, Torben G., Oleg Bondarenko, Albert S. Kyle, and Anna A. Obizhaeva. 2016. "Intraday Trading Invariance in the E-mini S&P 500 Futures Market." Working Paper, Available at SSRN 2693810.

Bae, Kyounghun, Albert S. Kyle, Eun Jung Lee, and Anna A. Obizhaeva. 2014. "An Invariance Relationship in the Number of Buy-Sell Switching Points." Working Paper, University of Maryland.

Black, Fischer. 1986. "Noise." *The Journal of Finance*, 41: 529Ű–543.

Grinold, Richard C. 1989. "The Fundamental Law of Active Management." *Journal of Portfolio Management*, 15(3): 30–37.

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70(3): 393Ű–408.

Kyle, Albert S. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies*, 56: 317–355.

Kyle, Albert S., and Anna A. Obizhaeva. 2016. "Market Microstructure Invariance: Empirical Hypothesis." *Econometrica*, 84(4): 1345–1404.

Kyle, Albert S., and Anna A. Obizhaeva. 2017*a*. "Dimensional Analysis and Market Microstructure Invariance." Working Paper. Available at SSRN 2785559.

Kyle, Albert S., and Anna A. Obizhaeva. 2017*b*. "Market Microstructure Invariance: A Dynamic Equilibrium Model." Working Paper. Available at SSRN 2749531.

Kyle, Albert S., Anna A. Obizhaeva, and Tugkan Tuzun. 2016. "Microstructure Invariance in U.S. Stock Market Trades." FEDS Working Paper No. 2016-034. Available at SSRN 2774039.

Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang. 2017. "Smooth Trading with Overconfidence and Market Power." *Review of Economic Studies*, Posted March 8: http://www.restud.com/paper/smooth–trading–with–overconfidence–and–market–power/.

Kyle, Albert S., Anna A. Obizhaeva, Nitish R. Sinha, and Tugkan Tuzun. 2014. “News Articles and the Invariance Hypothesis.” Working Paper, University of Maryland.

Murphy, Kevin M., Andrei Shleifer, and Robert W. Vishny. 1991. “The Allocation of Talent: Implications for Growth.” *The Quarterly Journal of Economics,* 106(2): 503–530.

Philippon, Thomas, and Ariell Reshef. 2012. “Wages and Human Capital in the U.S. Finance Industry: 1909Ű2006.” *The Quarterly Journal of Economics,* 127(4): 1551–1609.

Schwarzkopf, Yonathan, and Doyne J. Farmer. 2010. “Empirical Study of the Tails of Mutual Fund Size.” *Physical Review E,* 81: 066113.

Treynor, Jack. 1995. “The Only Game in Town.” *The Financial Analysts Journal,* 51(1): 81–83 (reprinted from The Financial Analysts Journal 22, 1971, 12–Ű14, 22).

Appendix

Proof of Theorem 1: Substituting equation (22) into equation (17) yields trader n's optimal demand (26). Substituting equation (26) into equation 9 yields the equilibrium price (27).

The second-order condition has the correct sign if and only if $\frac{2}{(N-1)\gamma} + \frac{A}{\tau} > 0$. Given the definition $\Delta := \frac{1}{\theta} - 2 - 2/(N-2)$, this is equivalent to

$$\frac{A}{\tau} \cdot \frac{N}{N-2} \cdot \frac{1}{\theta} \cdot \frac{1}{\Delta} > 0. \tag{A-1}$$

Therefore, assuming $N > 2$, the second-order condition holds if and only if $\Delta > 0$.

Proof of Lemma 1: If $m = n$, then from equations (2) and (3), we have

$$\tilde{i}_n - \tilde{i}_{-n} = \tau_H^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v}) + \tilde{e}_n - \frac{1}{N-1} \sum_{j \neq n} \theta \cdot \tau_H^{1/2} \cdot (\tau_v^{1/2} \tilde{v}) - \frac{1}{N-1} \sum_{j \neq n} \tilde{e}_j. \tag{A-2}$$

This yields $\mathrm{Var}_0^n[\tilde{i}_n - \tilde{i}_{-n}] = \tau_H(1-\theta)^2 + 1 + (N-1)^{-1}$.

If $m \neq n$, then

$$\tilde{i}_m - \tilde{i}_{-m} = \theta \cdot \tau_H^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v}) + \tilde{e}_m - \frac{1}{N-1} \left(\tau_H^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v}) + \tilde{e}_n + \theta \cdot \tau_H^{1/2} \cdot (\tau_v^{1/2} \cdot \tilde{v})(N-2) + \sum_{j \neq n, j \neq m} \tilde{e}_j \right). \tag{A-3}$$

This implies $\mathrm{Var}_0^n[\tilde{i}_m - \tilde{i}_{-m}] = \tau_H(1-\theta)^2 \cdot (N-1)^{-2} + 1 + (N-1)^{-1}$.

Proof of Theorem 3: Substituting $W = \hat{W} \cdot \tau_H^{-1/2}$ and $V = \hat{V} \cdot \tau_H^{-1/2}$ into equation (50), we get equation (67) in Theorem 3. Equation (51) yileds

$$\frac{\tilde{x} \cdot (\mathrm{Var}_0[p])^{1/2}}{N^{1/2}} \approx c_I \cdot \theta \cdot (\tilde{i}_n - \tilde{i}_{-n}) = \bar{c}_I \cdot \tau_H \cdot \theta \cdot (\tilde{i}_n - \tilde{i}_{-n}). \tag{A-4}$$

Equation (A-4) implies that

$$\tilde{I} := \frac{(\tau_H^{-1/2} \cdot \tilde{x}) \cdot \sigma_p}{(\tau_H \cdot N)^{1/2}} \approx \bar{c}_I \cdot \theta \cdot (\tilde{i}_n - \tilde{i}_{-n}). \tag{A-5}$$