PRELIMINARY

The value of information flows in the stock market

H. Duong [*] B. Taub[†][‡]

November 11, 2024

Abstract

Stock market traders who trade because of information they possess reveal that information to the rest of the market in the process of bidding: if the information is positive they bid up the price, and if it is negative they lower it. New information constantly develops and is brought to the market in this way, and because it influences prices, it ultimately influences the allocation of investments by firms.

Using a new approach, we estimate the flow of this information and the price of that information (different from the stock price), and thus the total value of that information, for each stock, and then sum up this value across all stocks, obtaining an estimate of the total value of the dynamic flow of information in the stock market as a whole. This requires digesting the records of tens of thousands of stock orders (including cancelled orders, not just executed trades) to construct the dynamic limit order book and estimate the information flow and value from its structure.

Our results support the notion that the cross-correlation of price impact across stocks is consistent with the CAPM: there is a single systematic component of price impact, and this is driven by the volatility of the systematic component of the stock market. This result suggests that by separating the underlying information into two components, systematic and idiosyncratic, informed traders distinguish be-

[*] University of Glasgow.

[†] University of Glasgow. This research was carried out in part during my stay at the European University Institute, Florence, with the support of a Fernand Braudel Fellowship. I am also grateful to AlgoSeek Corporation for generously donating the data that has been used in this study. I acknowledge with gratitude an allocation of computational resources from the XSEDE and ACCESS programs under the aegis of the National Science Foundation. We thank Iacopo Mastromatteo for a useful comment on an earlier draft.

[‡] output/November 11, 2024

tween productive assets that have a systematic impact on the economy and those that can be diversified.

1 Introduction

Capital must be allocated to myriad investments, properly balancing the relative risks and returns associated with each investment; this is done in stock markets, with information about each investment constantly flowing in and being weighed against information about competing investments.

This problem is important because the tens of thousands of individual traders in stock markets who have information have the incentive to keep that information private and to profit from it in trading, and it is important to understand whether their activities result in the efficient and proper allocation of investment resources when interacting in stock markets, as this determines the growth of the economy.

The literature has proposed a theory, the Kyle [1985] model, that shows how the equilibrium behavior of traders processing this information results in a relationship between the information, the price of assets, and the volatilities of the prices and trade volume of equities traded in the market. A fundamental quantity of interest in the Kyle [1985] model, λ, the slope of the supply curve reflects the fundamental forces driving the stock value; λ reflects the marginal effect of trading on the price, and so is known as the price impact parameter.

While Kyle [1985] predicts how λ is determined and quantitatively links it to the volatilities of price and trading volume, it was developed within the context of a dealership market without any reference to the limit order book. Furthermore, previous studies have not examined the relationship between λ and the limit order book, typically estimating price impact using execution data rather than LOB data. By drawing an analogy between λ and the slope of the order book, we observe that the limit order book has a visible structure: the set of unexecuted orders forms a pattern—essentially a supply curve—with a slope that is driven by the underlying incentives created by the information possessed by some of the traders.

In contrast to the traditional interpretation of Kyle's λ, which assumes its estimation originates from execution data, we propose an innovative approach by hypothesizing that Kyle's λ is expressed within the order book. Furthermore, we estimated price and volume volatilities through a different methodology focused on execution orders. While the price impact and volatility estimates are carried out using entirely different data and methods, they confirm the predictions of the Kyle model.

To test the hypothesis that price impact, Kyle's λ, is embodied in the limit order book, we run a simple regression of price against quantity in each updated instance of the limit order book to recover an estimate of λ, and relate it to our estimates of the volatilities. But because of the immense number of messages and the hundreds of trading days for each stock this is in itself a huge task. Notably, previous studies estimating λ from execution data did not incorporate limit order book data, making our data approach both novel and distinctive.

We carried out our analysis using a data set containing the complete set of transactions over a three-year period.[1] Order information is typically in extremely raw form: each order submitted by traders to buy or sell shares of stocks is recorded as a message in the so-called limit order book (LOB), with the stock ticker label, price, quantity, time of initiation, and time of execution or cancellation, and these orders and terminations occur on a rolling basis.

To carry out the analysis, the limit order book needs to be retrospectively reconstructed from this data, whilst handling the asynchrony of orders, that is, looking ahead and backward to link messages with their subsequent execution or cancellation. For a typical stock on a typical day, the number of messages will be in the tens of thousands, and for a few actively traded stocks, it is hundreds of thousands.

It is an important detail of market trading that the vast majority of messages—orders generated by computerized trading algorithms that are placed in the limit order book, waiting for a counterparty to agree to the trade a pre-specified price and quantity—are cancelled by those same algorithms before they are executed, often lasting just a few milliseconds. Using a small sample of 82 stocks over a one-month period, this starkly comes to the fore: 97 percent of all orders are cancelled before being filled. Theory predicts this; to test the theory it is essential to have the full limit order book data, as executed trades are only a small part of the story.

A key facet of the Kyle model, as outlined in Boulatov and Taub [2014], is that in addition to the original interpretation of λ as a measure of price impact, the inverse of λ, $1/\lambda$, is a Lagrange multiplier for the constraint characterizing the how information is dynamically resolved for the optimization problem solved by traders with private information. It is, therefore, a shadow price, and when expressed in conjunction with the constraint to which it is associated, it is the shadow price of information.

[1] We thank AlgoSeek corporation for generously donating this data.

Information has a precise definition in the Kyle model: it is the forecast error variance of the traders in the market who are not in possession of private information, and who must therefore glean information from the flow of orders that come to them in the market; these traders are designated as "market makers" in the Kyle model. Information is thus the ignorance of these market makers.

Combining the notion of information and a notion of the price of information, one can state the value of the information flowing into the market. This is our central aim, examined under both single-asset and multi-asset Kyle models.

Multiple assets

Multi-asset extensions of the Kyle model show how informed traders use information about multiple assets to trade a specific asset, using information they have about the correlation of the fundamental characteristics of these assets. Uninformed traders—market-makers in the terminology of the Kyle model—are aware of the informed traders' use of correlation information and price stocks with this in mind. This literature includes papers by Caballe and Krishnan [1994], Hasbrouck and Seppi [2001], Back, Cao, and Willard [2000], Seiler and Taub [2008] and Bernhardt and Taub [2008].

The optimal strategies in these theoretical models boil down to a matrix of coefficients that are applied to the signals that traders observe: in the case of informed traders, the matrix is of trading intensities: for each stock, they choose an amount to trade based on the direct observation of the true value of the stock, but their trade is augmented by correlated information from the direct observation of other stocks. The pricing coefficient, Λ, similarly takes a matrix form, using information from order flow from correlated stocks when pricing each individual stock. In each instance of the literature, however, the cross-asset correlation structure was left abstract and unstructured.

There is a completely different theory, the capital asset pricing model (CAPM), which accounts for correlation—the correlation of returns—across assets. The central conclusion of the CAPM is that in equilibrium, the correlation across assets is due to a single factor, systematic risk, with all other returns being idiosyncratic and uncorrelated across all individual assets and thus fully diversifiable. It, therefore, makes sense to conjecture that a similar division can be made when characterizing the cross-asset correlation of the fundamentals of stocks in the Kyle model; if this conjecture is correct, then all cross-asset correlation would boil down to common systematic risk.

If one could separate the systematic and idiosyncratic influences in the Kyle model, then it would be an immediate prediction that idiosyncratic shocks to fundamental asset values would be of no use in cross-asset trades, and so any optimal cross-asset trading strategies would reduce to diagonal matrix, and this would also be reflected in pricing that is, the equilibrium matrix describing Λ—the multi-asset version of Kyle's price impact measure λ—would be diagonal.

This is what we find. Our results support the notion that the cross-correlation of price impact across stocks is consistent with the CAPM: there is a single systematic component of price impact, and this is driven by the systematic component as captured by the volatility of the systematic component of the stock market.

The nature of private information

The information in the Kyle model is the forecast error variance of the uninformed market makers. This information can be measured, and its value can also be measured; we provide the estimate of this value, which, when normalized, is consistent across stocks.

What is the purpose of the information flow? Our results confirming the single source of correlation, systematic risk, suggest that by separating the underlying information into two components, systematic and idiosyncratic, informed traders distinguish between productive assets that have a systematic impact on the economy and those that can be diversified. From a CAPM perspective, this is the only information that matters, as any non-systematic value can be diversified away.

The structure of this paper is as follows. We first review relevant findings on the price impact (Section 2.2) and establish the framework for analyzing our order book event data (Section 2.3). We also relate the Kyle framework to dimensional analysis and invariance. Then, we present the details of the data and computation methods employed in our analysis (Section 2.4). Section 2.5 examines the price impact on the limit order book under both single-asset and multi-asset Kyle models. In Section 2.6, we explore the relationship between price impact and the value of information. The final section investigates the relationship between price impact and business cycles.

2 Related literature

There is a large literature focusing on price impact and, equivalently, liquidity. Amihud [2002] used daily and monthly trading data of stocks traded in the New York Stock Exchange (NYSE) in the years 1963–1997 to examine the effect of illiquidity. In his paper, he measured the illiquidity ratios as the time-series average of the daily ratios of the absolute value of percentage returns to dollar volume and found out that expected stock returns are an increasing function of expected illiquidity. Alternatively, following Hasbrouck [2009] and Goyenko, Holden, and Trzcinka [2009], a representative coefficient is estimated as the λ coefficient in the regression of the root square of dollar volume against the price. They show that their estimation and effective cost are moderately positively correlated. Bisias, Flood, Lo, and Valavanis [2012] estimates this measure daily by using all transactions during normal trading hours on each day. The authors estimate λ as the coefficient of regression of the natural logarithm of volume in dollars against the sequence of intraday returns.

Our study contributes to a growing body of literature on cross-asset price impact. The related theoretical work includes papers by Caballe and Krishnan [1994], Back et al. [2000], Seiler and Taub [2008] and Bernhardt and Taub [2008]. The multi-asset Kyle model was first studied by Caballe and Krishnan [1994] under the general setting with n assets and m informed traders. The generality of the model makes only a partial analysis of the solution possible. Back et al. [2000] considers the univariate Kyle model with n informed traders with correlated signals. They find that when signals of informed traders are perfectly correlated, there is no linear equilibrium. Bernhardt and Taub [2008] presents a one-period model of n risk-neutral informed traders and m assets. They allow informed traders to internalize how their trades impact the prices and trades of other speculators. They show that the covariance structure of asset fundamentals is the driver of prices, while the covariance of liquidity trade drives that of order flows. Seiler and Taub [2008] extend the analysis of Bernhardt and Taub [2008] to an infinite horizon model in which informed investors receive private long-lived information repeatedly.

A number of empirical studies show evidence of significant cross-price impact in stock markets, including Hasbrouck and Seppi [2001], Pasquariello and Vega [2015], Wang, Schafer, and Guhr [2016], Garcia del Molino, Mastromatteo, Benzaquen, and Bouchaud [2020], Mehdi Tomas and Benzaquen [2022]. Hasbrouck and Seppi [2001] decompose multi-asset order flows and returns, and find that two-thirds of the commonality in returns can be explained by commonality in order flows. Pasquariello and Vega [2015] investigate the trading activity in the New York Stock Exchange (NYSE) and the National Association of Securities Dealers Automated Quotation System (NASDAQ) stocks between 1993 and 2004 found that the cross-price impact is often negative and both direct and absolute cross-price impact are smaller when there are many speculators. Wang et al. [2016] use the intraday data of AAPL, GS, and XOM from the NASDAQ stock market in January 2008 and empirically show that cross-asset price impacts are small and appear to be transient instead of permanent. They also find the cross-correlation of the trade signs has a short memory.

Similar to our approach, Garcia del Molino et al. [2020] uses the multiple asset Kyle framework to estimate different price impact measures and shows that the Kyle estimator performs better in the market with heterogeneous volatility. Methodologically, our paper is closely related to papers that decompose the information in stock prices based on the CAPM, such as Hasbrouck and Seppi [2001]. However, unlike the prior papers that study information decomposition of stock returns, our paper aims to partition price impacts into two components: systematic influence and idiosyncratic influence. We find that the idiosyncratic shocks to fundamental asset values have little impact on cross-trades, and any optimal cross-asset trading strategies would reduce the trading intensity to a diagonal matrix.

Finally, this paper is consistent with an extensive body of literature studying the value of information flow and order flow in the stock market. Berk and van Binsbergen [2015] uses the Center for Research in Security Prices (CRSP) survivorship bias-free database of mutual funds to study the skill in the mutual fund industry. They find that each mutual fund has used its skills to generate about \$3.2 million per year and in the aggregate, mutual funds in the US markets made over \$19 billion per year. Yang and Zhu [2019] presents a model of the strategic interaction between fundamental investors and back-runners. They calibrate their model and estimate that the potential institutional investors' daily profits in the U.S. equity market are in the order of \$150 million per day. The results of these papers are in line with our empirical result as the institutional investors' profit is somewhat equivalent to the value of information. Even though our results are in the same order of magnitude as these estimations, we take a different approach. We extract the value of information separately for each stock from price and quote data, then find normalized information flow averaged across trading days.

3 LOB approach to price impact

Econophysics has gained traction by asserting that price impact is a consequence of executions and that this impact is inherently nonlinear. This view diverges sharply from traditional finance, primarily because econophysicists approach price impact in a purely phenomenological way, largely omitting the role of information and equilibrium from their models. **?** do empirical research on thousands of trades and point out that the autocorrelation of trade sides decays extremely slowly with time, and the price fluctuation is persistent and predictable. Therefore, they argue that price impact should not be linear and permanent. **?** demonstrate that order flow is autocorrelated: trades often cluster in the same direction, creating herding behaviors among traders. This autocorrelation magnifies price impact and contributes to increased volatility and substantial deviations from price equilibria assumed in classical models.

Another school of thought is the mainstream finance literature on market microstructure—which views price impact with greater emphasis on informational trading, equilibrium, transaction costs, and arbitrage opportunities. **?** shows that price impact is not uniform across assets or time periods. He considers price impact as a dual indicator: temporary price impact (associated with liquidity costs) and permanent price impact (associated with informational effects). **?** introduced innovative econometric techniques to isolate the impact of informed trading.

A primary model in this traditional literature is the Kyle [1985] model, which assumes that total order flow, from the perspective of a market maker, represents pure noise without price impact. However, the Kyle model also predicts that price impact will emerge from the perspective of the informed trader, whose filtration of information allows them to predict and benefit from subsequent price movements. This assumption would require econophysicists to argue for an ex-post understanding of this informational filtration—something often unaddressed in their phenomenological framework.

The Kyle model also makes clearer predictions about how price impact manifests. Under the Kyle framework, price impact is in the minds of market makers, yet it is realized in the LOB's structure, specifically in the slope. By measuring the LOB's slope, we can quantify price impact, expecting it to exhibit stationarity and even constancy if underlying variances remain fixed and persistent. Despite shifts in the LOB, the price impact should appear linear across different timestamps. Additionally, this slope will likely differ across various tickers, reflecting how price impact can vary by asset characteristics.

The hypothesis that price impact is expressed in the LOB rather than purely in transaction data (TAQ), presents a substantial econometric challenge. The sheer volume of data in the LOB vastly exceeds that in transaction logs, as approximately 97% of all LOB orders are ultimately canceled without execution. The Kyle model and the hypothesis that λ is realized in the LOB suggest that orders get canceled as new information arrives or as prices shift. This reflects traders' adjustments based on updated information or perceived changes in asset value. This abundance of order activity in the LOB, combined with the relative sparsity of executed trades, requires sophisticated statistical techniques to parse meaningful price impact data.

3.1 The elementary Kyle model

In this section, we analyze the basic static Kyle model. The model has three ingredients: (i) the variance of the fundamental value per share of the security that is being traded, Σ, for a Gaussian distributed value and from which the realized value is drawn; (ii) the variance of the order flow of the so-called "noise" traders, σ^2; and (iii) the price impact of any trade on the price, λ, which is determined by rational traders—"market makers"—who are not informed about the true value of the security as determined by fundamentals, and who are in competition with each other. The realized value of the security is privately observed by a single trader who then exploits this information in his trade. The noise traders' trades are treated as entirely exogenous; that is, they do not react to observations about price in any way.

In equilibrium, the following relationship holds:

$$\lambda = \frac{1}{2}\frac{\sqrt{\Sigma}}{\sigma} \tag{1}$$

This formula concerns the underlying structure of the model, but Σ and σ^2 are not directly observable. Defining Σ_P and the variance of observable price and σ_V^2 as the volatility of executed transaction volume, the following proposition establishes that formula (1) can be restated in terms of observables:

Proposition 3.1

$$\lambda = \frac{1}{2}\frac{\sqrt{\Sigma}}{\sigma} = \frac{\sqrt{\Sigma_P}}{\sigma_V} \tag{2}$$

Proof: See Appendix A. □

A brief aside concerning dimensional analysis and invariance

In any model in which the stock price and trading volume are functions of the volatilities of prices and volumes, then they must satisfy a homogeneity property. This is an instance of Buckingham's theorem and is also reflected in the papers of Kyle and Obizhaeva [2016] and Obezhayeva and Kyle [2017].

The argument is straightforward. Suppose that we posit that the price impact of trades is as follows:

$$\lambda \equiv \frac{\Delta P}{\Delta V} = f(\sigma_P, \sigma_V)$$

Because the volatilities are constructed from the differences of the levels of price and volume, if we apply a multiplicative factor k to the price, as would occur for example, in a stock split, then the volatilities must also reflect this scale factor:

$$\frac{k\Delta P}{\Delta V} = f(k\sigma_P, \sigma_V)$$

Now, choose the scale factor to, in fact, equal the inverse of the volatility:

$$\frac{\frac{\Delta P}{\sigma_P}}{\Delta V} = f(1, \sigma_V)$$

The same argument holds for the volume:

$$\frac{\frac{\Delta P}{\sigma_P}}{\frac{\Delta V}{\sigma_V}} = f(1, 1)$$

Thus, any statistical test would reasonably construct the left-hand side, which we can think of as normalized price impact, and then look for a constant on the right-hand side (if the theory has no further ingredients beyond price impact and volatilities). In the Kyle model, the predicted right-hand side constant is 1 (applying Proposition 3.1 to the Kyle model).

An additional observation is that if the underlying true model is linear, which the log of Kyle's ratio is, then it will be entirely vacuous to obtain regression coefficients of $1/2$ and $-1/2$ (using logs of the variances on the right-hand side). The only relevant result in such a regression is the intercept term.

Does this reasoning, that is, that it is trivial that the estimated coefficients are $1/2$, apply to the limit order book? In the limit order book model, the estimated volatilities come from *executed* prices and volumes, whilst, in the limit order book, the price impact is derived from the shape of the limit order book and need not be driven by the pattern of executions; most of the orders in the limit order book messages are, in fact, never executed. There is no theoretical a priori reason to expect the price impact in the limit order book to follow the Kyle model structure.

4 Data and computational details

We collected data from several sources. Our primary data source is a proprietary database of US stocks that are trading on the NASDAQ exchange. The database contains message-level information of all stocks from 2016 to 2018. For each stock, there is a raw message file that contains all trading messages of one stock sent to the market at high speeds in milliseconds within a trading day. The file provides a comprehensive record of every trade and order book change of all stocks on the exchange; there are approximately 6,500 stocks in all.

The first step in the empirical analysis is the reconstruction of the limit order book, moment by moment. As the dataset records all events that led to state changes to the order book, we can reconstruct limit order book for any stock at each moment in the trading day and at full depth for the specified period. The comprehensive and full-depth level data allow us to analyze different characteristics of price impact and its relationship with limit events with high accuracy.[2]

The message file contains every arriving market and limit orders as well as cancellations and updates of one stock. The information of the message file has 9 data fields.

1. "Date" provides information regarding the trading day
2. "Timestamp" All entries have a timestamp of seconds after midnight with the precision of milliseconds.
3. "Ordernumber", each order has a unique ID; subsequent actions such as execution, deletion or partial execution are indexed by the same number. Zero reference orders correspond to a hidden limit market order.
4. "EventType" There are 11 types of market events in the data. Provided details in a table in the additional appendix, Appendix B.1]

[2] Thus far, the empirical literature in this field has been limited to the use of preconstructed LOB data such as Lobster with only a few layers of top-of-the-book information.

5. "Ticker" provides information regarding the trading stock
6. "Price" the price of the order
7. "Quantity" the quantity of the order
8. "MPID" provides information of Market Participant Identifier. This identifier is used by FINRA member firms to report trades.
9. "Exchange" There are two main exchanges, ARCA (the electronic order book of the NYSE) and NASDAQ. All entries detail which exchanges the order was sent to.

Generally speaking, the order ID corresponds to the unique order reference number, which we can use to differentiate messages. However, there are some exceptions that may affect our limit order reconstruction.

1. All messages classified as "trade bid" and "trade ask" have zero reference orders. Those are hidden market orders with full information for all other fields except the order number. As they are market orders, they don't affect our limit order reconstruction, but we need to take them into consideration when we look at executed orders.
2. For big stocks that are trading across trading platforms, there are some order IDs corresponding to multiple different orders sent to different trading venues. One example is the messages with order ID 6168348 (TSLA, 08 Feb 2018). Essentially, the ID corresponds to 2 separate messages sent to different exchanges. The first order was a bid order at 08:20:56, which was sent to NASDAQ, then eventually got executed and filled later. The second order was an ask order at 09:42:40, which was sent to ARCA, then deleted eventually. To differentiate those different orders with the same reference number, we can look at the exchange and nature of the order. First, these orders were sent to different exchanges. We can use trading venues to find out and group all related orders. Second, we can use the nature, such as the order type and price, to map out all related orders. For example, "Add bid" orders should have related orders of type "execute bid" and "fill bid"; "Add ask" orders should have related orders of type "execute ask" and "fill ask."

Second, we obtain the data for the stock directory with market cap, R^2_{CAPM}, β_{CAPM} and variance from the NYSE [3] and Zoonova [4].

[3] Stock directory, https://www.nyse.com/listings_directory/stock

[4] Stock market watch,https://www.zoonova.com/Home/Markets

Our sample data directory contains all active stocks during the period between 1 January 2021 and 31 December 2021. All stocks must meet three pre-screening criteria to be in the directory: (1) it is a common stock (2) it is active on the first and last day during the sampling period. Active stocks refer to any stocks with trading activity on public exchanges during the sampling period. Out of over 6,500 tickers, some stocks were not listed or did not exist as of February 2018, (3) it has NASDAQ as the primary listing exchange. After filtering out all duplicates and erroneous entries, we are left with 6,481 stocks. In our initial study, we obtained a sample from February 2018; there were 19 trading days in total for each stock, and each trading day had between approximately 10,000 to over 10,000,000 messages for one stock. The primary justification for selecting this period is that February 2018 was a calm month, falling outside the U.S. earnings season and unaffected by major macroeconomic events.Therefore, the input file size can reach the region of 20 GB for one ticker on each trading day, thus posing technical challenges in terms of computation and data storage. We employed stratified random sampling by partitioning all tickers into subpopulations. The sample stocks were chosen based on the following sampling characteristics: high R^2, low R^2, high β, low β, high market cap, low market cap and low variance, high variance. 12 tickers were randomly selected from each group, yielding an initial sample of 96 tickers. After removing duplicate entries, 82 unique tickers remained. The rationale behind this sampling method is that stocks have high variances in all those characteristics. Stratified random sampling allows us to effectively select stocks that represent a diverse range of groups. The statistical summary of those stocks is illustrated in Table 1.

	R-squared	Marketcap	Yearly Price Variance	Beta
Mean	0.1590	19,300,751,911	6.096	0.93
Standard Error	0.0300	6,811,346,968	1.335	0.18
Median	0.0299	353,644,000	0.970	0.85
Minimum	0.0001	23,198	0.063	(2.47)
Maximum	0.7253	343,970,000,000	35.490	4.81

Table 1: Descriptive statistics of the sample

5 Testing Kyle model

This section empirically tests the Kyle model framework using the LOB data introduced in Section 2.4. We also discuss the relationship between price impact, volatilities, and the slope of the order book within both the static single-asset model and the multi-asset model.

5.1 Testing the univariate static model

To carry out tests of the model, we estimated three fundamental quantities: λ, σ_P and σ_V.

We estimate λ by calculating the slope of the LOB, yielding an estimate $\hat{\lambda}$, using a sample of 82 stocks for the 19 trading days in February 2018. The algorithm reconstructs the sequence of limit order books by parsing and sorting the raw file of messages for the day's trades for a single stock. Each trading message results in an update of the limit order book; each updated limit order book is called a snapshot. Each stock typically has tens of thousands of messages, so consequently, there are tens of thousands of snapshots; for the most heavily traded stocks, there are hundreds of thousands of snapshots.

We denote the collection of snapshots a ticker-day. These snapshots for each ticker-day are then statistically analyzed, with three key estimates being generated. First, the slope of each ticker is estimated with an OLS regression; we separately estimate the bid side (downward-sloping demand curve) and the ask side (upward-sloping supply curve); theory predicts that these slopes should be the same in absolute value. We re-estimate λ for each snapshot using OLS, compiling a list of estimated $\hat{\lambda}$s for later averaging. We ignore the spread at the top of the book. Figure 1 depicts the estimated λs (back lines) of an example of 4 tickers APDN, PSA, PZZA, THS on 01 September 2018 for both ask and bid sides. The blue bands are 95% confidence intervals of estimated λs. The yellow lines are price volatilities during the trading day, estimated as price quadratic variations by minute. Among all stocks, estimated λs are higher at the beginning and the end of the trading day. The main reason behind this trend is that at the beginning of the trading daily, the market makers start making the market, and at the end of the trading day, all market participants cancel their resting orders. Therefore, the LOB liquidity is low, and the price impact is bigger. If excluding the first half hour of the trading day and the last half hour of the trading day, the estimated λs are highly stable for all stocks.

Second, we estimate the variances of executed price and of executed paper3/APDN01022018.png

[5pt] AP

order flow. Estimating the volatilities is challenging because trades occur at random times. We compute $(\Delta P_t)^2$ for each interval, normalize by dividing by Δt for that interval, and then take the moving average to estimate each variance.

Empirical tests

Our first test implements the notion that invariance holds, even though any price impact that appears in the limit order book need not follow the invariance requirement. If invariance holds, then as we articulated in (3.1), the normalized price impact must equal a constant; for the Kyle model specifically, the normalized price impact is equal to 1. In our first test of the model, conjecturing that the Kyle invariance relationship holds in the limit order book, not just in executed prices, we calculated the normalized price impact ratio $\frac{\sigma_P}{\sigma_V}/\lambda$ using our λ estimates from the slope of the order book and the volatility estimates, obtaining values of .7 (bid side) and 1.1 (ask side) from a sample of about 82 tickers.

Comparisons with direct price impact

Conceptually, price impact concerns the impact of *executed* orders on price, that is, direct price impact, as this is a central concern of real traders. It is conceivable that direct price impact is driven by structure outside of the ken of the Kyle model. It, therefore, behooves us to compare the direct price impact with the impact we have measured in the limit order book.

In order to measure direct price impact we calculated the ratio $\frac{P_t - P_{t-1}}{y_t - y_{t-1}}$ for each instance t of an executed trade, and calculated the average for each ticker for our February 2018 sample. The resulting value, calculated in the same way as in the discussion of invariance, is 2.04, double that of the ratio using the slope of the order book to estimate price impact.

In scrutinizing this result further, we found that the results were widely scattered: some tickers had a direct price impact as high as 80 times as high as the LOB estimate of the ratio, some were a fraction, but the average ratio was 11.3 times higher in our sample. It is apparent that the tickers with the highest discrepancy were thinly traded, while for heavily traded stocks, the direct price impact and the LOB estimated λs were essentially the same.

To test this observation, we used market capitalization as a proxy for trading intensity, and indeed, we found that low-capitalization stocks have higher direct-versus-LOB price impact ratios. We also measured trading intensity directly, as the volume of executed order flow per minute for our sample of stocks, with identical results: high-order flow stocks had significantly lower executed price impact. In a regression of the ratio of executed price impact to LOB estimated λ versus the log of executed order flow per minute. the estimated coefficient for the dependent variable is -1.1 ($R^2 : .18$, t-statistic: -4.4.)

Our explanation is as follows. For thinly traded stocks, executions occur only infrequently relative to LOB order messages; therefore, for thinly traded stocks, there is significant LOB activity in between executions. The LOB activity reflects up-to-date information, whilst executions happen after a long evolution of information; therefore, execution is more likely to reflect new private information, and pricing reflects this.

More specifically, consider the LOB at times t, $t+1$, and $t+2$. At times t, the LOB has a specific slope, driven by the dictates of the private information volatilities underlying the fundamentals of the stock. At $t+1$ new information arrives and, if it is positive information, moves the *entire* LOB up; however, reflecting the thin trading, no execution takes place. At $t+2$ there is an execution—for the sake of discussion, a buy. The starting point for this execution is the new top of the book, which has moved up due to the cumulative arrival of information at time $t+1$ and also at $t+2$, but then, in addition, the order walks up the book. The walk up the book has the impact of the λ from the slope of the book, but the effect of the prior movement of the entire book is added to the impact, thus seeming to magnify it. The *cumulative* effect of the earlier arrival of the new information is combined with the book walk—so the price impact is bigger. The effect is bigger for thinly traded stocks because there are longer delays between executed trades. A technical argument is provided in Appendix C.

Regression tests

Our next statistical test consists of a simple linear regression of the logarithm of the averaged $\hat{\lambda}$ on the logarithmic transform of the formula (2) for 82 stocks, with the estimates of λ, Σ_P and σ_V averaged over 19 trading days in February 2018, treating each ticker as an observation. This yielded the results in Table 2:

The predicted values of the coefficients are 0 for the intercept term, $\frac{1}{2}$ and $-\frac{1}{2}$ respectively for the price volatility and volume volatility; basic statistical theory suggests that when explanatory variables are measured with error, as $\hat{\lambda}$, must be, the estimated coefficients are biased toward zero; the coefficients here thus reflect this bias; the intercept term is less successful; however, the P-value is weak. These results thus strongly support the Kyle

$R^2 = .88$ N=82 F: 315.7	Predicted value	Coefficient	Standard error	t-statistic	P-value
Intercept	0	1.13	0.41	2.75	.007
$\ln(\Sigma)$	$\frac{1}{2}$	.57	.023	24.56	9.61E-39
$\ln(\sigma^2)$	$-\frac{1}{2}$	-.336	.032	-10.40	1.83E-16

Table 2: Basic univariate static model regression results. (82 stocks, 19 trading days 2018)

model formulation and, more generally, an informational interpretation of stock market trading.

Given the panel structure of the dataset, a more robust approach than averaging λ over 19 trading days is to apply panel regression analysis. To determine whether fixed or random effects are more appropriate, a Hausman test is conducted, with the null hypothesis favoring the random effects model over the alternative, fixed effects model. The resulting p-value of 5.8×10^{-16} strongly suggests that the fixed effects model is more consistent. Additionally, to assess the presence of time effects, we perform a Lagrange Multiplier Test, which yields a p-value of 0.008416, indicating significant time effects.

We then estimated the model with both individual ticker and time effects, and the results are presented in Table 3

$R^2 = .891$ N =82 T = 19 F: 5965.7	Predicted value	Coefficient	Standard error	t-statistic	P-value
$\ln(\Sigma)$	$\frac{1}{2}$	.523	.0055	94.518	2.2E-16 ***
$\ln(\sigma^2)$	$-\frac{1}{2}$	-.2267	.00132	-22.141	2.2E-16 ***

Table 3: Static model panel regression results. (82 stocks, 19 trading days 2018)

In comparison to the previous model, the coefficient of $\ln(\Sigma)$ shows a slight decrease, approaching the theoretical value of $\frac{1}{2}$, while the coefficient of $\ln(\sigma^2)$ increases to -0.23. Overall, the estimated coefficients remain statistically significant and closely align with the predicted values. These findings provide further support for the validity of the Kyle model framework

5.2 Cross-asset correlation and the CAPM

As discussed in the introduction, a number of theoretical extensions of the Kyle model explore cross-asset effects. This literature does not ascribe the cross-asset correlations to particular causes, however the logic of the CAPM would point to a single cause of correlation, with stocks otherwise uncorrelated.

The CAPM perspective thus leads to a sharp prediction: that the cross-asset effects in the pricing matrix Λ are driven only by systematic factors. If there is a way to filter out these systematic factors, then the residual Λ matrix should be diagonal. We test this idea in two distinct ways.

First method: extraction from regression

For the first test of the correlation structure we enhance the regression of the logarithm of $\hat{\lambda}$ on the logarithms of the volatilities as in Table 2 above by including the CAPM R^2, displayed in Table 4:

$R^2 = .89$ N=82 F: 223.70	Predicted value	Coefficient	Standard error	t-statistic	P-value
Intercept	0	0.44	0.50	0.88	0.38
$\ln(\Sigma)$	$\frac{1}{2}$	0.53	0.03	17.65	5.62E-29
$\ln(\sigma^2)$	$-\frac{1}{2}$	-0.34	0.03	-10.78	3.96E-8
CAPM R^2		0.9	0.39	2.30	.024

Table 4: Basic univariate static model including CAPM R^2.

These results are still in accord with the underlying model in the sense that the coefficients on the volatility terms are consistent with the values of $\frac{1}{2}$ and $-\frac{1}{2}$ predicted by theory; the CAPM R^2 coefficient is statistically significant.

Second method: covariance matrices

The second method uses cross-asset information. As shown in Bernhardt and Taub [2008], one can express the equilibrium matrix Λ in terms of the cross-asset price and volume covariance matrices:

$$\Lambda = \Sigma^{1/2} \cdot \sigma^{-1}$$

where $\Sigma^{1/2}$ and σ are the Cholesky factors for the cross-asset price and volume covariance matrices, respectively; we can generalize this as with Proposition 3.1, we can demonstrate that the relationship holds for the covariance matrices of price and executed volume.[5]

One of the assets included in the portfolio of assets (again, the 19 trading days in February 2018) is SPY, the index fund tracking the S&P500, which is a widely accepted proxy for the systematic asset. We eliminate the rows and columns corresponding to SPY from the price and covariance matrices and calculate the resulting Λ for the remaining tickers. The resulting estimate of Λ should, in principle, have the influence of SPY removed and, if the CAPM intuition is correct, be driven solely by heterogenous fundamentals across the remaining assets. If the CAPM intuition is correct, the resulting residual matrix should be diagonal, as cross-asset information is irrelevant. The diagonal of the matrix is then the proper estimate of the Λ associated with idiosyncratic firm value.

Norm comparisons

If the hypothesis that the cross-correlation between the λ values is due entirely to correlation with the systematic market process, then when we remove the SPY rows and columns from the Λ matrix, the matrix should be essentially diagonal, that is, each stock's λ value should not be affected by any other stock, other than SPY. Therefore, the matrix norm of the reduced Λ matrix should be driven solely by the diagonal. Carrying out this experiment using the trace norm yields a sum of the absolute values of the eigenvalues of the idiosyncratic matrix of 0.0395, whereas the similar sum with the diagonal removed is 0.000545, that is, essentially zero.

Alternatively, we can compare the matrix norms of the two matrices, where the matrix norm is the maximum singular value; in this case, the norms with and without the diagonals are 0.021576 and 0.0153497 respectively, again demonstrating that the off-diagonal correlation is reduced relative to the diagonal. These calculations support the hypothesis of CAPM-driven correlation.

Comparing the two methods

One can roughly calculate the correlation of the idiosyncratic λ values calculated using the first method and the second method. We carried this out

[5] The derivations for the two-asset version of the model are set out in Appendix D.

with the sample of 82 tickers, again limited to the 19 trading days in February 2018. Using the first method, we calculated a predicted-λ series in which the effect of the R^2 term was dropped; intuitively, this series would roughly capture the non-systematic element of the variances. We then calculated the correlation between the forecasted idiosyncratic series with the diagonal of the Λ matrix with the SPY elements removed. The correlation between these two measures is .52; that is, there is a significant degree of correlation, suggesting that both methods at least partially succeed in isolating and extracting the idiosyncratic component of Λ.

The conclusion we draw is that, using two different approaches, that there appears to be a single factor driving cross-asset correlation, and that the two approaches yield measures of the correlation that are very closely correlated.

6 Measuring information flows

New information is constantly brought to the market. Traders keep the information private to preserve their advantage. Nevertheless, traders impart their information, and prices reflect it after trading. The Kyle model quantifies this process. In Kyle's framework, price impact is a result the incorporation of private information into asset prices.

Can this information be *measured*? Yes, it is the market makers' *forecast error variance* Σ_t. It is related to the variance of price.

Does the information have a price, and can it be measured? Yes: it is related to λ: the shadow price is $1/\lambda$. Boulatov and Taub [2014] provides the theoretical justification for price interpretation: $1/\lambda$ is the Lagrange multiplier for the constraint facing the informed trader, expressing how the market maker's forecast error variance decays as a result of trade, with the "income" in the constraint equal to the forecast error variance. Therefore, we can use the estimates of λ, and also the price volatility estimates, to estimate the quantity and price of information.

Using the Boulatov-Taub interpretation of the inverse $1/\lambda$ as the shadow value of information and the variance of price Σ as a proxy for the market makers' forecast error variance as a measure of the information, one can then calculate the value of information on a per-share basis as the product of these two quantities.

Using the basic structure of the Kyle model, this information value flow can be shown to be equivalent to the profit for the informed trader, which is equal to the product of the volatilities of price and order flow.

$$\frac{\Sigma}{\lambda} = \sqrt{\Sigma}\sigma$$

Using the estimates of λ and volatilities for 82 tickers in February 2018 from the previous section, the correlation between these two measures is 0.944.

Normalized information flow

Using 82 tickers over the month of February 2018, we can analyze information flow value. The information flow is per unit of value for each ticker; thus, by dividing the value of information by the value of shares traded (price times volume, each averaged over the trading day), one obtains the normalized information flow; this average value is about 0.024, and the median value is of 0.0021, but with a wide variance. Another way to calculate the average normalized information value is to divide the total value of information of all stocks by the total trading volumes. Using this method, we yield the normalized information flow of 7.5×10^{-5}. For the longitudinal data for Wednesday trading over three years, 2016, 2017, and 2018, the estimated normalized information flow values are 1.6×10^{-4}, 1.6×10^{-4}, and 1.79×10^{-4} respectively.

Define this information flow parameter as ω. We can multiply the normalized flow by the value of all stocks traded to obtain an estimate of the value of all information flowing in the economy, divided into systematic and idiosyncratic elements. Multiplying the total daily Nasdaq trading value of about \$300 billion [6] by the normalized information flow $\omega \sim .00016$, you obtain \$48 million per day. This is in close accord with the estimate of Yang and Zhu [2019] as discussed in the introduction.

7 Business cycle effects

Given the support for the CAPM-driven correlation hypothesis in section 2.5, a natural conjecture is that the systematic component of λ might be correlated with the business cycle. This influence could be driven by changes in the systematic part of the volatility of fundamental asset values, that is, volatility of returns seems to rise in recessions.

To test this hypothesis, we calculated the value of information flows over the business cycle. We again used a sample of stocks for every Wednesday

[6] Nasdaq, https://www.nasdaqtrader.com/Trader.aspx?id=DailyMarketSummary

spanning the three years 2016-2018. After winnowing the sample to exclude tickers that did not span the whole period,[7] out of an initial sample of 49 tickers, this left a sample of 29 tickers.

We then carried out the similar exercise of estimating the slope of the bid side of the LOB for each ticker on each day by averaging the calculated OLS slopes for each snapshot and also estimating the price volatility and executed-volume volatility for each ticker-day. We then calculated the normalized information value flow, $\frac{\frac{\Sigma_P}{\lambda}}{PV}$ that is, the value of information flow relative to the average value of trade for that ticker and day. This yields a dimensionless constant. We regressed this constant against the volatility of the SPY.[8] The coefficients, t-statistics, R^2, and P-values were then averaged. The results are displayed in Table 5.

	Coefficient	t-statistic	$\lvert t\text{-statistic}\rvert$
	estimate	average	
Intercept average	8.60×10^{-4}	1.8	2.78
|Intercept| average	1.34×10^{-3}	4.06	4.1
Σ_{SPY} average	8.86×10^{-2}		
R^2 average	.12		

Table 5: Normalized information flow versus Σ_{SPY}. (Mean(Σ_{SPY}) = .0078)

The average of the absolute value of the t-statistics is included to reflect the fact that many of the estimated intercept terms in the regressions are negative, whereas most of the estimated coefficients of the SPY volatility are positive. Thus, there appears to be a systematic component of normalized information flow value.

The magnitudes of the information flow value from the systematic part, which can be roughly estimated as the product of the coefficient on Σ_{SPY} with the mean of Σ_{SPY}, yields $0.0886 \times .0078 = 6.91 \times 10^{-4}$; the magnitude from the absolute values of the idiosyncratic parts are 1.34×10^{-3}, which is of similar magnitude. The average of the signed values of the intercept terms is 8.6×10^{-4}, an even smaller number.

We can conclude that the value of information increases during times of high systematic return volatility, that is, during recessions, as reflected in

[7] This potentially biases the results due to survivorship biases.

[8] The volatility of the SPY is similar in spirit to the VIX volatility index for the S&P500 index. However, the VIX is the volatility of the *return* to the S&P500 index, whereas the SPY volatility is the volatility of the *level* of the index.

the countercyclicality of the VIX and the SPY volatility.

Information flow and CAPM variables

Given the correlation of the normalized information flow with the volatility of the SPY (analog of the VIX), is there a relationship with the CAPM? We regressed the normalized information value, averaged over the three years for each ticker, against the CAPM beta and the CAPM R^2 for each ticker. The result is in Table 6. (The regressions on the CAPM beta did not yield significant results.)

R^2: .22	Coefficient value	t-statistic	P-value
Intercept	6.2×10^{-4}	0.97	.337
CAPM R^2	0.000976	2.055	0.049

Table 6: Normalized information flow (three-year average, each ticker) versus CAPM R^2.

The results strongly support the hypothesis that systematic information flows are strongly correlated with the business cycle, as high-R^2 stocks have higher information flows.

There is an additional conclusion: because, like the VIX, the Σ_{SPY} is strongly correlated with the business cycle, it is a proxy for the underlying systematic process. CAPM reasoning suggests that investors and traders care only about the systematic component of the value of any stock. The statistical significance of the coefficient on the Σ_{SPY} suggests that the only information that matters for traders for any stock is the systematic component, and the value of the information is the value of unearthing and isolating the information about the systematic part of the information.

8 Conclusion

By treating the Kyle model's price impact parameter, λ, as the slope of the limit order book, our results strongly support the validity of the Kyle [1985] model.

While the theory literature has developed a number of models analyzing how cross-asset correlation of the underlying fundamental asset values influences the cross-asset correlation of the corresponding price impacts, our results support the notion that the cross-correlation of price impact across stocks is consistent with the CAPM: there is a single systematic component of price impact, and this is driven by the systematic component as captured by the volatility of the systematic component of the stock market, that is, the SPY volatility, and this systematic component of the underlying value is responsible for any cross-asset correlation, and any concomitant correlation of the price impact measures, that is, the λs.

The information in the Kyle model is the forecast error variance of the uninformed market makers. This information can be measured, and its value can also be measured. When normalized by the value of trade in each ticker, the value of the information flow per unit of value is on the order of .0005. This number accords well with the overall income of firms engaging in stock market trading.

The normalized information flow value is strongly countercyclical, that is, it is strongly correlated with the volatility of the overall market. The connection of the information flow with the SPY volatility is strongly confirmed by the strong correlation of the normalized information flow, averaged over time, with the degree to which the stock is influenced by the aggregate market, that is, the CAPM R^2 value of the stock.

What is the purpose of the information flow? By separating the underlying information into two components, systematic and idiosyncratic, the traders distinguish between productive assets that have a systematic impact on the economy and those that can be diversified. From a CAPM perspective, this is the only information that matters, as any non-systematic value can be diversified away.

A Proofs and derivations

Proof of Proposition 3.1. Notation:

V	underlying value of asset
x	informed trader's trade $= \beta V$
u	"noise" trade
y	total order flow $x + u$
β	informed trader's trading intensity with solution $\frac{1}{2\lambda}$

Proof:

$$\begin{aligned} \operatorname{var}(y) &= \operatorname{var}(x+u) = \operatorname{var}(x) + \operatorname{var}(u) \\ &= \beta^2 \Sigma + \sigma^2 = \frac{1}{4\lambda^2}\Sigma + \sigma^2 \\ &= \frac{4\sigma^2}{4\Sigma}\Sigma + \sigma^2 = 2\sigma^2 \end{aligned}$$

Also,

$$\operatorname{var}(P) = \operatorname{var}(\lambda y) = \lambda^2 2\sigma^2 = \frac{\Sigma}{4\sigma^2}2\sigma^2 = \frac{1}{2}\Sigma$$

$$\rightarrow \sqrt{\frac{\text{var(executed price)}}{\text{var(executed volume)}}} = \sqrt{\frac{\frac{1}{2}\Sigma}{2\sigma^2}} = \frac{1}{2}\frac{\sqrt{\Sigma}}{\sigma} = \lambda$$

□

B Details of the message data

B.1 Data structure

Twelve types of market events are recorded in the data (see Table 6 below). As the limit order book is a primary focus, "cross" messages that occurred in a dark pool or an auction are filtered out. When a market order is matched against several limit orders, each matching is recorded separately. Messages labeled as "FILL ASK" and "FILL BID" have missing price and quantity fields. We need to trace back to the original order of the same IDs to figure out the missing pieces.

Event Type	Description
ADD ASK	Submit a new ask order
ADD BID	Submit a new bid order
CANCEL BID	Cancel the bid order partly
CANCEL ASK	Cancel the ask order partly
CROSS	Dark pool transactions without price and quantity
DELETE ASK	Delete the whole ask order
DELETE BID	Delete the whole ask order
EXECUTE ASK	Execute the order partly
EXECUTE BID	Execute the order partly
FILL ASK	Fill the ask order completely
FILL BID	Fill the bid order completely
TRADE BID	Fill the bid order completely

Table 7: Even Type in the message file

To reconstruct a limit order book from a raw message file we follow the following procedure.

1. **Step 1**. Eliminate abnormal messages that aren't with the active region. As we observed, messages with a price above or under 1.5 times the average price were normally not within the active region. We filtered out those messages out of the sample. We also handle missing data from "FILL ASK" and "FILL BID" as mentioned above.

2. **Step 2** Construct a first snapshot with only the first order book event.

3. **Step 3** Iterate over all new events to construct all snapshots and store them in an array. The newly constructed limit order book snapshot has a full depth with price and volume at all levels. For "ADD

ASK" and "ASK BID" message types, a new snapshot is updated by adding those new messages to the previous snapshot. For the "CANCEL BID", "CANCEL ASK", "EXECUTE ASK", and "EXECUTE BID" message types, the order ID and exchange of the message are matched against the orders in the previous snapshot to look for the outstanding order that should be updated by reducing its order size. For the "DELETE BID", "DELETE ASK", "FILL ASK", and "FILL BID" message types, the corresponding orders get processed completely. Therefore, the new snapshot is constructed by deleting all orders with the same IDs of incoming messages. At any time, there are only "ADD ASK" and "ADD BID" messages outstanding in a snapshot. Upon the creation of a snapshot, ask and bid order types are separated, sorted, and grouped by price. The final step is to filter out abnormal entries and then create the cumulative depths at each price level.

To illustrate the above procedure, we assume that the initial snapshot of a stock (ACN) has 2 outstanding orders as follows.

Timestamp	OrderNumber	EventType	Ticker	Price	Quantity	Exchange
00:00.0	120	ADD BID	ACN	60.02	50	ARCA
00:00.0	129	ADD ASK	ACN	68	4	ARCA

In the next period, a new "ADD BID" message of 40 shares arrives at the price of 61, the snapshot will be updated by adding the new message to the new snapshot.

Timestamp	OrderNumber	EventType	Ticker	Price	Quantity	Exchange
00:00.0	120	ADD BID	ACN	60.02	50	ARCA
00:00.0	129	ADD ASK	ACN	174.7	4	ARCA
00:00.0	138	ADD BID	ACN	62	40	ARCA

Right after, the trader of order ID 120 wants to reduce his order size, so he submits a 'CANCEL BID" message of 20 shares. The new snapshot will updated by reducing his outstanding order size by 20 shares.

B.2 Sample covariance matrices

For the univariate model, we use quadratic variations for volume and price variance to capture all market variances. For multi-asset models, each asset

Timestamp	OrderNumber	EventType	Ticker	Price	Quantity	Exchange
00:00.0	120	ADD BID	ACN	60.02	30	ARCA
00:00.0	129	ADD ASK	ACN	174.7	4	ARCA
00:00.0	138	ADD BID	ACN	62	40	ARCA

has a different execution pattern and time frame. In order to calculate the quadratic covariance matrices for price and volume, we divide the trading into a uniform grid in time $t_0, ..., t_n$ with a timescale $t_k - t_{k-1} = 600$ seconds. In this way, price and volume changes of all assets have the same dimensions. The price at t_k (p_{t_k})is defined as the executed price of the closet execution order before t_k. The volume at t_k (v_{t_k})is defined as the cumulative volume between t_{k-1} and t_k.

C Direct price impact

We assume that between 2 executed orders at times $t_0 = 0, t_{n+1} = T$, there are n limit order book events at $0 < t_1 < ... < t_n < T$ with the volume ΔV_{t_i}. The price change between 0 and T can be defined as

$$\Delta p = p_T - p_0 = \sum_{i=1}^{n+1} (p_{t_i} - p_{t_{i-1}}) = \sum_{i=1}^{n+1} \Delta p_{t_i} \tag{3}$$

Where p_{t_i} is the contribution components of events i to the direct price impact.

In the case of no execution order, we can interpret the Δp_{t_i} as the change of the shadow price or the change in the fundamental values of the asset because of the arrival of the limit order event at t_i. If the event at t_i is an executed order, we have $\Delta p_{t_i} = \Delta V_{t_i} \lambda$. If the event at t_i is not an executed order, we define $\Delta p_{t_i} = \alpha_{t_i} \Delta V_{t_i} \lambda$. The reason for this definition is that an Δp_{t_i} increasing function of ΔV_{t_i}. For example, a market maker should react more strongly to the big order at the top of the book. Second, we can justify this assumption by considering α_{t_i} as a function of the probability of execution. Another way is to interpret α_{t_i} as a discount factor of the information content of the order. If we substitute those equations into the equation (3) and obtain.

$$\frac{p_T - p_0}{\Delta V_T} = \sum_{i=1}^{n} \alpha_{t_i} \frac{\Delta V_{t_i}}{V_T} \lambda + \lambda = (\sum_{i=1}^{n} \alpha_{t_i} \frac{\Delta V_{t_i}}{V_T} + 1) \lambda \tag{4}$$

If we define $a_{t_i} = \alpha_{t_i} \frac{\Delta V_{t_i}}{V_T}$, we can rearrange and arrive at the following expression.

$$\frac{\frac{p_T - p_0}{\Delta V_T}}{\lambda} = (\sum_{i=1}^{n} a_{t_i} + 1) \tag{5}$$

The absolute value of the right-hand side of the above equation is greater than 1 if $\sum_{i=1}^{n} a_{t_i} \geq 0$. If we assume the limit order book events are symmetric. For any limit order book event type on the ask side, there is a corresponding type on the bid side. For example, “add bids” and “add ask” are a corresponding pair. The effects of these 2 corresponding orders on the price impact are exactly opposite. If the assumption that the limit order events are symmetric holds, on average $a_i = -a_{-i}$ for all $(i, -i)$ which are corresponding event types. In other words, the average of $\frac{\frac{p_T - p_0}{\Delta V_T}}{\lambda} = 1$. Therefore, the effects of all limit order events converge to the direct price impact for a sufficiently long time.

D Two-asset correlation model

Bernhardt and Taub [2008] sets out a static model of a multi-asset Kyle model in which asset values are cross-correlated, exploring how informed speculators with differential information about the spectrum of assets exploit that information in trading correlated assets. Informed speculators use cross-asset information to trade strategically if they can observe prices. If prices are unobserved before trade, they do not use the information.

The purpose of this note is to translate the cross-asset speculation models (CAM) into a slightly simpler setting in which there are just two assets, one of which is the systematic asset (in practice, the S&P 500 index fund, SPY), and the other of which is an ordinary stock with positive correlation driven by CAPM considerations.

Main derivations

There are N informed traders and M assets. In the basic model of interest, $N = 1$ and $M = 2$. The value of asset 1 is

$$\begin{aligned} v_1 &= v_{11}e_1 \\ v_2 &= v_{21}e_1 + v_{22}e_2 \end{aligned} \tag{6}$$

so

$$v = Ve, \qquad V \equiv \begin{pmatrix} v_{11} & 0 \\ v_{21} & v_{22} \end{pmatrix} \tag{7}$$

Thus, v_1 is the systematic asset value, and v_2 is the heterogeneous asset; moreover with this interpretation,

$$v_{11} = 1 \tag{8}$$

To maintain the spirit of the basic static Kyle model, we can assume that there is only a single informed speculator, and so the signal structure is

$$\begin{pmatrix} s_1^1 \\ s_2^1 \end{pmatrix} = \begin{pmatrix} A_{11}^1 & A_{12}^1 \\ A_{21}^1 & A_{22}^1 \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{9}$$

Because it is a CAPM-driven model we can assume that the informed trader has full information about the systematic asset, that is,

$$A = \begin{pmatrix} 1 & 0 \\ A_{21}^1 & A_{22}^1 \end{pmatrix} \tag{10}$$

which implies that he has full information about the second asset as well after netting out the systematic part, leaving

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{11}$$

The trading strategy for the informed trader is

$$\begin{aligned} x_1^1 &= b_{11}^1 s_1^1 + b_{12}^1 s_2^1 + B_{11}^1(X_1 + u_1) + B_{12}^1(X_2 + u_2) \\ x_2^1 &= b_{21}^1 s_1^1 + b_{22}^1 s_2^1 + B_{21}^1(X_1 + u_1) + B_{22}^1(X_2 + u_2) \end{aligned} \tag{12}$$

or

$$\begin{pmatrix} x_1^1 \\ x_2^1 \end{pmatrix} = \begin{pmatrix} b_{11}^1 & b_{12}^1 \\ b_{21}^1 & b_{22}^1 \end{pmatrix} \begin{pmatrix} s_1^1 \\ s_2^1 \end{pmatrix} + \begin{pmatrix} B_{11}^1 & B_{12}^1 \\ B_{21}^1 & B_{22}^1 \end{pmatrix} \begin{pmatrix} X_1 + u_1 \\ X_2 + u_2 \end{pmatrix} \tag{13}$$

The pricing rule is given by

$$\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} \begin{pmatrix} X_1 + u_1 \\ X_2 + u_2 \end{pmatrix} \tag{14}$$

Define

$$\Gamma \equiv I + \sum_{k=1}^{N} \gamma^k$$

Recalling that N is the number of informed traders, in the basic Kyle model $N = 1$; here we are assuming two assets, so $M = 2$.

Equilibrium formulas

Defining the total order flow covariance matrix

$$\Psi \equiv \begin{pmatrix} bA & I \end{pmatrix} \begin{pmatrix} \Sigma_e & 0 \\ 0 & \Sigma_u \end{pmatrix} \begin{pmatrix} A'b' \\ I \end{pmatrix} \tag{15}$$

and skipping to Proposition 2 in Bernhardt and Taub [2008], we have

$$\begin{aligned} b^i &= A^1 \Sigma_e V'(I + \gamma^i) \\ \gamma^{i\prime} &= -\Psi^{-1} bA\Sigma_e A^{i\prime} b^{i\prime} \\ \Gamma'\lambda' &= \Psi^{-1} bA\Sigma_e V' \end{aligned} \tag{16}$$

where Σ_e is the variance-covariance matrix of the fundamentals, and V is the vector of realized asset fundamental values.

Next we can state Proposition 4, which provides a formula for the direct trading intensities:

$$b \sim (A\Sigma_e A')^{-1/2}\Sigma_u^{1/2} \tag{17}$$

Notice that this reduces to the fundamental static Kyle model formula if there is one trader and one asset,

$$b = \frac{\sigma_u}{\Sigma_0^{1/2}} \tag{18}$$

However the key measurable quantity is λ:

$$\lambda = V\Sigma_e A b' \Psi^{-1\prime}\Gamma^{-1} \tag{19}$$

The quantities on the right hand side need to be related to the observables, namely price and total order flow. First, substituting from (17),

$$\lambda \sim V\Sigma_e A \Sigma_u^{1/2}(A\Sigma_e A')^{-1/2}\Psi^{-1\prime}\Gamma^{-1} \tag{20}$$

(noting the "$\sim$" rather than "=").We can also substitute from (17) into (15):

$$\Psi \sim \begin{pmatrix}(A\Sigma_e A')^{-1/2}\Sigma_u^{1/2}A & I\end{pmatrix}\begin{pmatrix}\Sigma_e & 0\\ 0 & \Sigma_u\end{pmatrix}\begin{pmatrix}A'\Sigma_u^{1/2}(A\Sigma_e A')^{-1/2}\\ I\end{pmatrix} \tag{21}$$

so that

$$\lambda \sim V\Sigma_e A\Sigma_u^{1/2}(A\Sigma_e A')^{-1/2}\left(\begin{pmatrix}(A\Sigma_e A')^{-1/2}\Sigma_u^{1/2}A & I\end{pmatrix}\begin{pmatrix}\Sigma_e & 0\\ 0 & \Sigma_u\end{pmatrix}\begin{pmatrix}A'\Sigma_u^{1/2}(A\Sigma_e A')^{-1/2}\\ I\end{pmatrix}\right)^{-1}\Gamma^{-1} \tag{22}$$

Also, we can reduce Γ:

$$\begin{aligned}\Gamma = I + \sum_{k=1}^{N}\gamma^k &= I - \sum_{k=1}^{N} bA\Sigma_e A^{i\prime}b^{i\prime}\Psi^{-1}\\ &= I - \sum_{k=1}^{N}(A\Sigma_e A')^{-1/2}\Sigma_u^{1/2}A\Sigma_e A^{i\prime}b^{i\prime}\left(\begin{pmatrix}\Sigma_e & 0\\ 0 & \Sigma_u\end{pmatrix}\begin{pmatrix}A'\Sigma_u^{1/2}(A\Sigma_e A')^{-1/2}\\ I\end{pmatrix}\right)^{-1}\end{aligned} \tag{23}$$

where b^i has been left unreduced.

The covariance matrix of prices is as follows:

$$\lambda\Gamma\Psi\Gamma'\lambda' \tag{24}$$

Because Ψ is the covariance matrix of total order flow, we can in principal recover $\lambda\Gamma$ by factoring the observed Ψ, and also calculating the price covariance matrix. The tricky part is Γ. From Proposition 7, the covariance matrix of total order flow is $\Sigma_u\Gamma'$ [See proof of Proposition 7, p. 41 of Bernhardt and Taub [2008].]

Multiple-asset trading

The logic of the cross-asset paper Bernhardt and Taub [2008] presupposes an environment in which there are just a few relevant stocks. However there are thousands of stocks, and to capture the appropriate $\lambda\Gamma$ for a single stock i, we would need to add up the cross-asset $\lambda_{ij}\Gamma$ for all stocks j. What makes more sense is to treat SPY as the cross-asset ticker and isolate the effect of SPY on the $\lambda\Gamma$ of each of the smaller stocks.

By subtracting the influence of SPY, we can isolate the trade on private information unique to each ticker. This information, and also its value, can then be added up.

E Converting CAPM returns to prices

The paper Boulatov and Taub [2014] sets out a dynamic version of the Kyle model in which there are multiple stocks, the underlying value of which, and also the prices, can be correlated. There is a completely separate literature on the correlation across stocks, the CAPM, but this is a theory of stock *returns*, not prices. The purpose of this note is to demonstrate that one can compute the correlations of stock prices if one is given the βs of the stocks, and importantly, also the $\mathbf{R}^2$ attached to the stock by the CAPM structure.

In the CAPM the correlation across stocks is driven entirely by the market return, which they share, as the residuals in the CAPM return equation are inherently mutually independent. The magnitude of the correlation is then determined by the βs and the $\mathbf{R}^2$s, but the magnitude requires some calculations, which are presented here.

The calculations use an approximation result as a key step, and this approximation result is outlined in Appendix F.

Main derivations

From the CAPM we have the following characterization of the *returns* for asset i:

$$R_t^i = r + \beta^i(R_t^M - r) + e_t^i \tag{25}$$

We want to convert this equation into an equation relating the *prices* to the aggregate prices. Begin by taking logs:

$$\ln\left(\frac{P_t^i}{P_{t-1}^i}\right) = r + \beta^i\left(\ln\left(\frac{S_t}{S_{t-1}}\right) - r\right) + e_t^i \tag{26}$$

It is worth noting that even though e_t^i and e_t^j are independent, they don't necessarily have the same variance, that is, we need to keep in mind that $\sigma_{ei}^2 \neq \sigma_{ej}^2$ is possible.

Taking the exponential yields

$$P_t^i = e^{(1-\beta_t^i)r+e_t^i}\left(\frac{S_t}{S_{t-1}}\right)^{\beta^i} P_{t-1}^i \tag{27}$$

The market price itself is a process:

$$\frac{S_t}{S_{t-1}} = \frac{S_{t-1}+dS}{S_{t-1}} = 1 + R^M dt + \sigma^M dZ_t \tag{28}$$

where dZ_t is the systematic risk process. The price level equation becomes

$$P_t^i = e^{(1-\beta_t^i)r+e_t^i}\left(1+R^M dt+\sigma^M dZ_t\right)^{\beta^i} P_{t-1}^i \tag{29}$$

Expressing this in level terms yields

$$P_t^i = e^{(1-\beta_t^i)r+e_t^i}\left(1+R^M+\sigma^M \zeta_t\right)^{\beta^i} P_{t-1}^i \tag{30}$$

where ζ_t is the innovation of the systematic return process. This can now be decomposed into idiosyncratic and systematic parts. Taking logs,

$$\begin{aligned}\ln(P_t^i) &= \ln(P_{t-1}^i) + (1-\beta_t^i)r + e_t^i + \beta^i\left(1+R^M+\sigma^M\zeta_t\right)\\ &= \ln(P_{t-1}^i) + (1-\beta_t^i)r + \beta^i\left(1+R^M\right) + e_t^i + \beta^i\sigma^M\zeta_t\end{aligned} \tag{31}$$

However the systematic coefficient is a mixture of the idiosyncratic and systematic shocks.

The next question is how the decomposition of the shocks into idiosyncratic and systematic parts translates into the multi-asset Kyle model. But this has a known answer from the model in Seiler and Taub [2008]. That paper does not decompose the value shocks into idiosyncratic and systematic parts, however it does treat correlation across prices. The main issue however is the fact that equation (31) is in logs, whereas the model is in terms of levels. The correlation structure can however be calculated by using equation (30).

$$\operatorname{corr}\left(P_t^i, P_t^j\right) = \frac{\operatorname{cov}\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}, e^{e_t^j}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)}{\operatorname{var}\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)^{1/2}\operatorname{var}\left(e^{e_t^j}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)^{1/2}} \tag{32}$$

(Note that the terms P_{t-1}^i and P_{t-1}^j cancel in the correlation formula.) Thus, if we can estimate the CAPM elements for a particular stock ticker, and also estimate the systematic ζ_t process, then we can calculate the correlation and apply using the model from Boulatov and Taub [2014].

Thus, we can develop the correlation simply from the CAPM residuals for the tickers. The variances of the residuals can in turn be calculated from the R^2 statistics of the CAPM equations, which are given along with the β^i coefficients for each of the tickers. Specifically, we have

$$\operatorname{var} R_t^i = \operatorname{var}\left(r+\beta^i(R^M-r)+e^i\right) = \beta_i^2\sigma_M^2+\sigma_{e_i}^2 \to \mathbf{R}_i^2 = \frac{\beta_i^2\sigma_M^2}{\beta_i^2\sigma_M^2+\sigma_{e_i}^2} \tag{33}$$

Therefore

$$\sigma^2_{e_i} = \frac{\beta_i^2\sigma_M^2}{\mathbf{R}_i^2} - \beta_i^2\sigma_M^2 = \frac{(1-\mathbf{R}_i^2)}{\mathbf{R}_i^2}\beta_i^2\sigma_M^2 \tag{34}$$

and this can then be used to calculate var $\left(e^{e_t^j}\right)$. (Notice that because the β_i and $\mathbf{R}_i^2$ are different across tickers, the expected values can also differ.) Specifically,

$$E\left[e^{e_i}\right] = e^{\frac{1}{2}\sigma^2_{e_i}} \qquad \text{var}\left[e^{e_i}\right] = E\left[e^{2e_i}\right] - \left(E\left[e^{e_i}\right]\right)^2 = \left(e^{\sigma^2_{e_i}} - 1\right)e^{\sigma^2_{e_i}}$$

Similarly,

$$\text{cov}\left(e^{e_t^i}, e^{e_t^j}\right) = E\left[e^{e_i+e_j}\right] - \left(E\left[e^{e_i}\right]E\left[e^{e_i}\right]\right) = 0$$

To calculate the covariance we need to calculate the expected value of the product

$$E\left[\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\left(e^{e_t^j}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)\right]$$

and subtract the product of the expectations. The product of the expectations is direct from the calculations already done. The expected value of the product is

$$\begin{aligned}
&E\left[\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\left(e^{e_t^j}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)\right]\\
=&E\left[e^{e_t^i}e^{e_t^j}\right]E\left[\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)\right]\\
=&E\left[e^{e_t^i}\right]E\left[e^{e_t^j}\right]E\left[\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)\right]\\
=&\left[e^{\frac{1}{2}\sigma^2_{e_i}}e^{\frac{1}{2}\sigma^2_{e_j}}\right]E\left[\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right)\right]\\
\approx&\left[e^{\frac{1}{2}\left(\sigma^2_{e_i}+\sigma^2_{e_j}\right)}\right]e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma^2_M}
\end{aligned}$$

where the first equality follows from the independence of the e_i from ζ, and the second equality comes from the independence of the e_i, and finally the approximation result from the appendix is used.

The product of the expectations is more straightforward:

$$\begin{aligned}
&E\left[\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)\right]\\
=&\left[e^{\frac{1}{2}\sigma^2}\right]E\left[\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right]\\
\approx&\left[e^{\frac{1}{2}\sigma^2}\right]\left[e^{\beta^iR^M+\frac{1}{2}\beta_i^2\sigma_M^2}\right]
\end{aligned}$$

Yielding the product

$$\begin{aligned}
&\left[e^{\frac{1}{2}\sigma_{e_i}^2}\right]\left[e^{\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2}\right]\left[e^{\frac{1}{2}\sigma_{e_j}^2}\right]\left[e^{\beta^j R^M+\frac{1}{2}\beta_j^2\sigma_M^2}\right]\\
=&\left[e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)}\right]\left[e^{(\beta_i+\beta_j)R^M+\frac{1}{2}(\beta_i^2+\beta_j^2)\sigma_M^2}\right]
\end{aligned}$$

The covariance is then the difference

$$\begin{aligned}
&\left[e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)}\right]e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}-\left[e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)}\right]\left[e^{(\beta^i+\beta^j)R^M+\frac{1}{2}(\beta_i^2+\beta_j^2)\sigma_M^2}\right]\\
=&e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)}e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}\left(1-e^{(-\beta_i\beta_j)\sigma_M^2}\right)\\
=&e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)}e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta_i^2+\beta_j^2)\sigma_M^2}\left(e^{(\beta_i\beta_j)\sigma_M^2}-1\right)
\end{aligned}$$

The variance is not a simple variation on the covariance. The expectation of the product is

$$\begin{aligned}
&E\left[\left(e^{e_t^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)^2\right]\\
=&E\left[e^{2e_t^i}\right]E\left[\left(\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right)^2\right]\\
\approx&e^{2\sigma_{e_i}^2}\left[e^{2\beta^i R^M+2\beta_i^2\sigma_M^2}\right]
\end{aligned}$$

The square of the expectation is

$$e^{\sigma_{e_i}^2}e^{2(\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2)}$$

So the variance is the difference

$$\begin{aligned}
&e^{2\sigma_{e_i}^2}e^{2\beta^i R^M+2\beta_i^2\sigma_M^2}-e^{\sigma_{e_i}^2}e^{2(\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2)}\\
=&e^{\sigma_{e_i}^2}e^{2\beta^i R^M}e^{\beta_i^2\sigma_M^2}\left(e^{\sigma_{e_i}^2}e^{\beta_i^2\sigma_M^2}-1\right)
\end{aligned}$$

Thus, there is an interaction between the two variances such that the variance does not cleave into two separate parts.

Combining to form the correlation, and using the approximation of the ratio for the terms involving ζ_t from the appendix, we have the reduced expression

$$
\begin{aligned}
\operatorname{corr}\left(P_t^i, P_t^j\right) &= \frac{e^{\frac{1}{2}\left(\sigma_{e_i}^2+\sigma_{e_j}^2\right)} e^{(\beta^i+\beta^j)R^M} e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}\left(1-e^{(-\beta_i\beta_j)\sigma_M^2}\right)}{\left(e^{\sigma_{e_i}^2} e^{2\beta^i R^M} e^{\beta_i^2\sigma_M^2}\left(e^{\sigma_{e_i}^2} e^{\beta_i^2\sigma_M^2}-1\right)\right)^{1/2}\left(e^{\sigma_{e_j}^2} e^{2\beta^j R^M} e^{\beta_j^2\sigma_M^2}\left(e^{\sigma_{e_j}^2} e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}} \\
&= \frac{e^{(\beta^i+\beta^j)R^M} e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}\left(1-e^{(-\beta_i\beta_j)\sigma_M^2}\right)}{\left(e^{2\beta^i R^M} e^{\beta_i^2\sigma_M^2}\left(e^{\sigma_{e_i}^2} e^{\beta_i^2\sigma_M^2}-1\right)\right)^{1/2}\left(e^{2\beta^j R^M} e^{\beta_j^2\sigma_M^2}\left(e^{\sigma_{e_j}^2} e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}} \\
&= \frac{e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}\left(1-e^{(-\beta_i\beta_j)\sigma_M^2}\right)}{\left(e^{\beta_i^2\sigma_M^2}\left(e^{\sigma_{e_i}^2} e^{\beta_i^2\sigma_M^2}-1\right)\right)^{1/2}\left(e^{\beta_j^2\sigma_M^2}\left(e^{\sigma_{e_j}^2} e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}} \\
&= \frac{e^{\frac{1}{2}(\beta_i^2+\beta_j^2)\sigma_M^2}\left(e^{(\beta_i\beta_j)\sigma_M^2}-1\right)}{\left(e^{\beta_i^2\sigma_M^2}\left(e^{\sigma_{e_i}^2} e^{\beta_i^2\sigma_M^2}-1\right)\right)^{1/2}\left(e^{\beta_j^2\sigma_M^2}\left(e^{\sigma_{e_j}^2} e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}} \\
&= \frac{\left(e^{(\beta_i\beta_j)\sigma_M^2}-1\right)}{\left(e^{\sigma_{e_i}^2} e^{\beta_i^2\sigma_M^2}-1\right)^{1/2}\left(e^{\sigma_{e_j}^2} e^{\beta_j^2\sigma_M^2}-1\right)^{1/2}} = \frac{\left(e^{(\beta_i\beta_j)\sigma_M^2}-1\right)}{\left(e^{\sigma_{e_i}^2+\beta_i^2\sigma_M^2}-1\right)^{1/2}\left(e^{\sigma_{e_j}^2+\beta_j^2\sigma_M^2}-1\right)^{1/2}}
\end{aligned}
\tag{35}
$$

where the last equality emphasizes that the $\sigma_{e_i}^2$ term is added in the exponent, not multiplied. Evidently the e_i terms reduce the correlation, which is intuitively sensible in that the idiosyncratic error reduces the effect of the systematic risk, equivalent to reducing the $\mathbf{R}^2$.

F Approximation

We want to compute

$$\text{var}\,\left[\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right]$$

Use an approximation:

$$\begin{aligned}
E\left[\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\right] &= E\left[e^{\beta^i\ln\left(1+R^M+\sigma_M\zeta_t\right)}\right]\\
&\approx E\left[e^{\beta^i\left(R^M+\sigma_M\zeta_t\right)}\right]\\
&= e^{\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2}
\end{aligned}$$

Thus the variance approximation is the expectation of the square minus the squared expectation:

$$e^{2\beta^i R^M+4\frac{1}{2}\beta_i^2\sigma_M^2}-e^{2\left(\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2\right)}=e^{2\beta^i R^M+\beta_i^2\sigma_M^2}\left(e^{\beta_i^2\sigma_M^2}-1\right)$$

The covariance approximation calculations will be similar. A reminder that

$$\text{cov}(x,y)=E[xy]-E[x]E[y]$$

Thus,

$$\begin{aligned}
E\left[\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^j}\right] &= E\left[e^{\beta^i\ln\left(1+R^M+\sigma_M\zeta_t\right)}e^{\beta^j\ln\left(1+R^M+\sigma_M\zeta_t\right)}\right]\\
&\approx E\left[e^{\beta^i\left(R^M+\sigma_M\zeta_t\right)}e^{\beta^j\left(R^M+\sigma_M\zeta_t\right)}\right]\\
&= e^{(\beta^i+\beta^j)R^M}E\left[e^{(\beta^i+\beta^j)\sigma_M\zeta_t}\right]\\
&= e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}
\end{aligned}$$

Thus,

$$\begin{aligned}
\text{cov}(x,y) &\approx e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}-e^{\beta^i R^M+\frac{1}{2}\beta_i^2\sigma_M^2}e^{\beta^j R^M+\frac{1}{2}\beta_j^2\sigma_M^2}\\
&= e^{(\beta^i+\beta^j)R^M}\left(e^{\frac{1}{2}(\beta^i+\beta^j)^2\sigma_M^2}-e^{\frac{1}{2}\left(\beta_i^2+\beta_j^2\right)\sigma_M^2}\right)\\
&= e^{(\beta^i+\beta^j)R^M}e^{\frac{1}{2}\left(\beta_i^2+\beta_j^2\right)\sigma_M^2}\left(e^{\beta^i\beta^j\sigma_M^2}-1\right)
\end{aligned}$$

The correlation ratio is then

$$\frac{e^{(\beta^i+\beta^j)R^M} e^{\frac{1}{2}\left(\beta_i^2+\beta_j^2\right)\sigma_M^2}\left(e^{\beta^i\beta^j\sigma_M^2}-1\right)}{\left(e^{2\beta^i R^M+\beta_i^2\sigma_M^2}\left(e^{\beta_i^2\sigma_M^2}-1\right)e^{2\beta^j R^M+\beta_j^2\sigma_M^2}\left(e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}}$$

$$=\frac{e^{\frac{1}{2}\left(\beta_i^2+\beta_j^2\right)\sigma_M^2}\left(e^{\beta^i\beta^j\sigma_M^2}-1\right)}{\left(e^{\beta_i^2\sigma_M^2}\left(e^{\beta_i^2\sigma_M^2}-1\right)e^{\beta_j^2\sigma_M^2}\left(e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}}$$

$$=\frac{\left(e^{\beta^i\beta^j\sigma_M^2}-1\right)}{\left(\left(e^{\beta_i^2\sigma_M^2}-1\right)\left(e^{\beta_j^2\sigma_M^2}-1\right)\right)^{1/2}}$$

Notice that this is equal to 1 if $\beta^i = \beta^j$.

A more precise calculation can be carried out using the Taylor series approximation of $\left(1+R^M+\sigma^M\zeta_t\right)^{\beta^i}$.

References

Yakov Amihud. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets, 5(1):31–56, 2002.

K. Back, H. Cao, and G. Willard. Price manipulation and quasi-arbitrage. Journal of Finance, 55:2117– 2155, 2000.

Jonathan B. Berk and Jules H. van Binsbergen. Measuring skill in the mutual fund industry. Journal of Financial Economics, 118(1):1–20, 2015.

D. Bernhardt and B. Taub. Cross-asset speculation in stock markets. Journal of Finance, 2008.

D. Bisias, M. Flood, A. Lo, and S. Valavanis. A survey of systemic risk analytics. Annual Review of Financial Economics, 4:255–296, 2012.

A. Boulatov and B. Taub. Liquidity and the marginal value of information. Economic Theory, 55:307–334, 2014.

J. Caballe and M. Krishnan. Imperfect competition in a multi-security market with risk neutrality,. Econometrica, 1994.

Luis Carlos Garcia del Molino, Iacopo Mastromatteo, Michael Benzaquen, and Jean-Philippe Bouchaud. The multivariate kyle model: More is different. SIAM Journal on Financial Mathematics, 11(2):327–357, 2020.

Ruslan Y. Goyenko, Craig W. Holden, and Charles A. Trzcinka. Do liquidity measures measure liquidity? Journal of Financial Economics, 92(2):153–181, 2009.

J. Hasbrouck and D. Seppi. Common factors in prices, order flows, and liquidity. Journal of Financial Economics, 59:383–411, 2001.

Joel Hasbrouck. Trading costs and returns for u.s. equities: Estimating effective costs from daily data. The Journal of Finance, 64(3):1445–1477, 2009.

A. Kyle. Continuous auctions and insider trading. Econometrica, 53:1315–1355, 1985.

Albert S Kyle and Anna A Obizhaeva. Market microstructure invariance: Empirical hypotheses. Econometrica, 84(4):1345–1404, 2016.

Iacopo Mastromatteo Mehdi Tomas and Michael Benzaquen. How to build a cross-impact model from first principles: theoretical requirements and empirical results. Quantitative Finance, 22(6):1017–1036, 2022.

A. Obezhayeva and A. Kyle. Dimensional analysis, leverage neutrality, and market microstructure invariance. Working Paper, New Economic School, 2017.

Paolo Pasquariello and Clara Vega. Strategic cross-trading in the u.s. stock market. Review of Finance, 19(1):229–282, 2015.

P. Seiler and B. Taub. The dynamics of strategic information flows in stock markets. Finance and Stochastics, 2008.

Shanshan Wang, Rudi Schafer, and Thomas Guhr. Cross-response in correlated financial markets: individual stocks. European Physical Journal, 89(4):1–16, 2016.

Liyan Yang and Haoxiang Zhu. Back-Running: Seeking and Hiding Fundamental Information in Order Flows*. The Review of Financial Studies, 33(4):1484–1533, 07 2019.