12-2023

How commonality persists? (Through investors' sentiment and attention)

Chyng Wen TEE
Singapore Management University, cwtee@smu.edu.sg

Raja VELU

Zhaoque ZHOU

How Commonality Persists?
(Through Investors' Sentiment and Attention)

Abstract

Studies on commonality generally attribute the variation in asset returns to the variation in order flows. In this research study, we show that order flows do not predict asset returns, rather their relationship have been static over time. Thus we model both returns and the order flows as endogenous variables, and use investors' sentiment and attention as exogenous factors via a reduced-rank regression. We provide empirical evidence to demonstrate that cross-sectional commonality in attention (sentiment) is linearly (nonlinearly) associated with both returns and order flows at the intraday level, while the sentiment and attention measures themselves exhibit a nonlinear mutual relationship, thus revealing the multi-dimensional complex aspect of the commonality relationship. The persistence of this relationship over a decade is documented using a large sample of assets. The concept of commonality is also related to portfolio optimization.

Keywords: *commonality; return and order flow measures; co-movement; market sentiment; investor attention; principal component analyses; canonical correlation analyses; reduced rank regression; quantitative trading strategies.*

1. Introduction

The proliferation of social media as the primary platform for information exchange among market participants, has led to a surge of research interest in studying the relationship between investors' social media behavior and stock market returns. These are predominantly driven by the way investors process stock-related information—the literature broadly categorizes the media information into investors' (1) attention and (2) sentiment.

Investors pay varying degree of "*attention*" to different stocks. Attention being a scarce cognitive resource, there is a tendency for investors to group stocks into broad categories and trade select stocks from these, instead of focusing on individual stocks on their own. Order flows will naturally be directed to stocks with a high level of investors' attention, while stocks with low a level of attention exhibit muted trading activity. On a related but separate note, investors attach different "*sentiment*" to stocks under their attention based on past experience and the information available. Stocks attributed to a positive sentiment have a tendency to rise in prices, and vice versa. Consequently, sentiment provides a means to estimate how far a stock price might deviate from its economic fundamentals, while attention provides a means to gauge trading activity. Though sentiment and attention represent two different factor dimensions that may influence the trading behavior, they are not necessarily orthogonal to each other.

Academic research has long established the role played by investors sentiment on stock returns. Barberis, Shleifer, and Wurgler (2005) has introduced the view that sentiment can delink returns comovement from fundamental value. Subsequent research by Baker and Wurgler (2006), Kumar and Lee (2006), Jiang, Lee, Martin, and Zhou (2019) have furnished further empirical evidence of investors' sentiment being an informative, albeit noisy, predictor for short-term stock returns. On a closely related topic, studies initiated by Peng and Xiong (2006), and subsequently extended by **?**, Chen, Tang, Yao, and Zhou (2022) and Jiang, Liu, Peng, and Wang (2022), have advocated the role played by investors' attention, which captures a different dimension of influence on investors behavior. More recently, intense research interest has focused on the role played by social media (see, for instance, Cookson, Lu, Mullins, and Niessner (2022) and Cookson, Engelberg, and Mullins (2023)) on both attention and sentiment.

The literature has largely focused on the level of attention that a specific firm receives. Most research concentrates on the comovement of attention and sentiment, the common factors they share, and their correlation with the stock returns, with principal component analysis (PCA) and multi-variate regression being the standard tools used for these analyses. An open research problem is the role of investors' sentiment and attention on the investors' cross-sectional behavior in short-horizon, their mutual dependence, and their relationship to the cross-sectional commonality in stock returns and order flows. In this paper, we formulate a unifying framework via reduced-rank regression that is ideally suited for this multi-dimensional study. Unlike other earlier studies (see Chordia, Roll, and Subrahmanyam (2000) and Hasbrouck and Seppi (2001)) where order flows are taken to be exogenous influencing the returns, we demonstrate that both returns and order flows should be considered as endogenous variables and in our microstructure model, we study how both are driven by exogenous stock-related behavioral variables, notably investors' attention and sentiment. Our postulate is also empirically supported by other studies, such as Han, Huang, Huang, and Zhou (2022).

The literature attributes the source of commonality to the comovement of the demand or supply of liquidity, and more recently due to trading activities. Below we list the sources of commonality noted already in the literature:

1. Market markers use shared capital, and the source of commonality is mainly due to covariation in shared stocks (see Coughenour and Saad (2004)).
2. Stocks that are part of an index share common features (see Harford and Kaul (2005)).
3. Stocks are grouped into price tiers and the stocks within a tier exhibit commonality (see Green and Hwang (2009)).
4. Commonality is driven by program traders who may have a similar reading of common signals (see Corwin and Lipson (2011)).
5. Common stock ownership across funds (seeKoch, Ruenzi, and Starks (2016)).
6. Increased presence of HFT (see Malceniece, Malcenieks, and Putniņš (2019)).

Our work extends on Drake, Jennings, Roulstone, and Thornock (2017), who studied comovement in attention, and showed that the comovement of investor attention has market

consequences. Our objective is to connect earlier studies on commonality between returns and order flows to the commonality in investors' behavior, measured through attention and sentiment. Our analyses reveal insightful aspects of their mutual relationship, shedding light on issues that cannot be fully explained when they are studied in isolation.

We set out to investigate the interactions of commonality, building on the foundation of prior work focusing on pairwise correlation between sentiment or attention *vis-à-vis* individual stock return. We are particularly interested in the relationship between sentiment and attention commonalities with respect to the return and order flow commonalities, and in possible variation due to some common components. Because these measures are generally held to contain informed components, can commonality in these measures account for the commonality of short-term returns and order flows? This is an important topic that extends the market microstructure theory to incorporate recent advancement in behavioral finance.

Our paper makes several key contribution to the literature. We assimilate the latest views and insights on the importance of investors' sentiment and attention measures into a unifying framework. Instead of considering them as alternative metrics, we demonstrate that the two measures capture different aspect of investors' behavioral characteristics, and they exhibit a nonlinear mutual relationship. Empirically, high attention is associated with strong (positive or negative) sentiment, while low attention is associated with weak (muted or neutral) sentiment. Following this reasoning, we find supporting evidence that the variation in sentiment commonality is higher (lower) for stocks under high (low) attention. This can be attributed to confirmatory bias, where herding mentality results in investors echoing sentiments similar to their own for stocks under high attention. Similarly, the variation in the attention commonality should be higher (lower) for stocks with neutral (strongly positive or negative) sentiment—strong sentiments are attention grabbing, while neutral sentiment leads to inattention instead.

We demonstrate that both measures can provide useful information on return and order flow, but the associative relationship fall in different dimensions. We test the hypothesis that return is associated with sentiment, while order flow is associated with attention. The relationship between return and order flow is driven by the association between sentiment and attention measures. Investors buy or sell a stock based on the sentiment they held on the stock, which

in turn leads to a positive or negative return. On the other hand, trading volume is driven primarily by investors' attention—stocks under high attention experience large trading volume, but inattention leads to low trading activity. Our results provide a strong empirical confirmation that the persistent nature in the way investors interpret market sentiment and attention as an important source of commonality. A major advantage of our proposed approach from gauging sentiment and attention based on news analytics is that this relationship can be exploited for the development of real-time trading strategies as well.

Our results presented in this paper connect two extensive strands of literature: (1) studies on commonality and its sources, and (2) studies on behavioral finance, notably investors' sentiment and attention, and its impact on the market. This paper is organized as follows: after a brief overview on the descriptive statistics in Section 2, we begin with an exposition of our modeling framework along with a description of the data set used in our study in Section 3, including the high-frequency Twitter feed used in our sentiment & attention analysis. A longitudinal study is presented in Section 4 to investigate the dynamics of commonality over an extended time period, while the analysis of the results on the sources of commonality is covered in Section 5. To demonstrate the application of our framework, we apply our model to formulate quantitative trading strategies in Section 6. Finally, conclusions are drawn in Section 7. The main finding of this paper is that both order flows and returns are endogenous and their joint variation is due to variation in investor's sentiment and attention. In addition to correlation analysis using a large scale higher frequency data, we demonstrate causality in the relationship between the endogenous variables and the exogenous variables whenever possible within the limitations of our data.

2. Descriptive Statistics and Hypothesis Development

2.1. Commonality in Return and Order Flow

Historically, market microstructure research has focused mostly in the setting of a single security at a time. Since the early 2000s, both academics and practitioners have paid more attention to the magnitudes of cross-sectional interactions among stocks at the micro level.

Asset pricing theory suggests that stocks returns are correlated with the market return in general, which implies that, stocks move in tandem with the "market factor" on average. Most studies in commonality focus on this market factor, either by regressing stock returns, order flows, or liquidity measures on an aggregated index, or by extracting an index through the first principal component of the correlation or the covariance matrix of returns.

There are two common methodologies used for studying commonality: regression analysis and Principal Component Analysis (PCA). In the regression approach, the characteristic of an individual asset, such as return, order flow, or a liquidity measure, is regressed on a market level composite metric, and the significance of the slope coefficient is used to infer commonality (see Chordia et al. (2000)). In the PCA approach, after standardization of the variables over individual assets, the principal components are extracted from the resulting correlation matrix. The percentage of total variance explained by the first few components is usually taken to be an indicator of commonality (see Hasbrouck and Seppi (2001)). The first component in returns is usually interpreted as representing the market factor because of the uniform distribution of coefficient loadings. It is also hypothesized that because order flows may contain informed components, they are likely to explain the commonality in returns. This analysis is typically done via canonical correlation analysis between returns and order flows, or between their respective principal components. The leading indices of returns and order flows are shown to be highly correlated.

Identifying the common features or co-movement of financial time series has been the focus of numerous well-known studies, including Engle and Kozicki (1993), Vahid and Engle (1993) among others. Given that it has been consistently demonstrated that the first component represents the market factor year after year, the persistence of commonality over time is distinguished through the second component onwards. As observed in Aït-Sahalia and Xiu (2019), the common variation in financial sector is shown to be the differentiating factor for the persistence between the normal and the volatile years. This is consistent with the observation '... the liquidity measures seem to be influenced by both a market and an industry component;' in Chordia et al. (2000).

Although most commonality studies are confined to the first factor, there is a considerable

interest to go beyond the first PC and study the multi-dimensional aspect of the commonality. Boehmer, Li, and Saar (2018) go beyond the first PC to differentiate various high-frequency trading strategies. For example, the top three PCs are identified as corresponding to cross-venue arbitrage strategy, market making, and shot-horizon directional speculation. Calomiris and Mamaysky (2019) use PCA to investigate how news and its context drive risk and returns. They report that the first PC tracks the aggregate time series of market sentiment, while the second PC indicates break at the timing of the global financial crisis. More recently, Langlois (2020) uses PCA to investigate the multi-dimensional aspect of (systematic and idiosyncratic) skewness of stock returns.

2.2. Investors' Sentiment and Attention

Although classical finance theory defines no role for investor sentiment or attention—assuming that market participants are rational investors—recent research on behavioral finance has nevertheless refuted this view. With the increasing speed of diffusion of information and the presence of algorithmic traders, commonality in investors' sentiment and attention can be taken as the exogenous factors driving the commonality in returns and order flows.

The relationship between investor sentiment and stock returns has been actively explored in finance. With the real time dissemination of data from electronic news media, research in the higher frequency setting is now possible. Kumar and Lee (2006) show how the collective sentiments of retail investors can lead to return co-movements. They examine this by studying the common directional component in the buy-sell trading activities, and how changes in the investor sentiment inferred from this component can impact the co-movement in stock returns. Retail investors' trades tend to be correlated, and therefore order (buy-sell) imbalance across non-overlapping portfolios is taken as an indicator of investors sentiment. The cross-sectional regression model used in their paper includes various control factors, including the Fama-French and momentum factors. It is shown that using lead-lag relationships, the sentiment measures generally exert impact on small stocks, on stocks with lower nominal prices, and on stocks with low institutional ownership.

In support of this hypothesis, several studies in behavioral finance have postulated that market sentiment can lead to price changes that are consistent with economic fundamentals. Baker and Wurgler (2006) propose an index based on the first principal component of six (now five) macro measures, namely 1) turnover, 2) closed-end fund discount, 3) initial public offering, 4) first day premium, 5) equity issues, and 6) dividend premium. Because the turnover measure was not stable, it has been excluded from the index in recent versions. The monthly index is shown to correlate with major stock market movements. The review paper by Zhou (2018) outlines other measures as well.

Along the same theme, the role of media coverage that disseminates financial information to a broad audience has been extensively studied in recent literature. Fang and Peress (2009) observe, based on a cross-sectional study, that stocks with high media coverage tend to earn less than stocks with no media coverage. This relationship is shown to hold true even after controlling for the widely-recognized risk factors. At the macro level, how news articles can predict the returns and volatilities have been well documented by Calomiris and Mamaysky (2019). But to extend this to the study of commonality at the high frequency level, we need to understand how sentiment is formed instantaneously through news, and also how it can impact the stock performance in higher frequencies. Recent research also explores how managers who have access to better information form their sentiment, and suggest measures that can affect their attention (see Jiang et al. (2019)). It is argued that managers sentiment can be different from investors sentiment, and they provide a different dimension to the behavioral aspect of investment decisions. In this paper, it is our intention to unify these two related streams by carefully parsing through the variables that make up these measures, and we also investigate how they could influence the return and order flow commonalities differently.

Jiang et al. (2019) summarize various sentiment measures starting with the consumer confidence indices released by the University of Michigan and Conference Board that have long been considered to be good barometers of future economic outlook. They suggest fourteen monthly economic variables that may be aligned with macroeconomic business cycles. Many of these variables are available at the stock level—for instance, the textual tone of the conference calls on financial reporting of the company performance serves as proxies to the managers' sentiment.

They report that predictive power of managers' sentiment is far greater than that of commonly used macroeconomic variables.

Huang, Huang, and Lin (2019) demonstrate that when there are major external events that grab the investors attention broadly, investors would try to focus more on learning about market-wide shocks than about firm-related shocks. Proxies for the attention measures have been widely investigated in the literature as well. A comprehensive list can be found in Chen et al. (2022), which includes abnormal trading volume (calculated as the ratio of trading volume to the average over the previous year), extreme returns, past monthly cumulative return over the prior year, analyst coverage of earnings per share forecasts, and changes in the advertising expenses. While these anomaly measures are available at the stock level, a market indicator is also suggested, measuring the nearness of prior year NYSE index to historical high. An attention index can be obtained either by equal weighing of these measures, or by PCA or Partial Least Squares methods to gauge the aggregated level of investor attention. Some of these attention measures are available on a continual basis, and can be used in a high-frequency context as well, which we will develop further in this paper.

Investors' attention can also influence commonality in stocks' return and order flows, an effect which we will study in this paper. In fact, attention measure can itself exhibit comovement, as demonstrated by Drake et al. (2017)—they study the extent to which investor attention to a firm is explained by attention paid to the firm's industry and the market in general. Cziraki, Mondria, and Wu (2021) argue that attention measure can also exhibit a location effect by showing that stocks earn higher returns when they attract abnormally high asymmetric attention from local investors.

Chu, He, Li, and Tu (2022) demonstrate that although fundamental variables can be strong predictors when sentiment is low, they tend to lose their predictive power when investor sentiment is high. Non-fundamental predictors perform well during high-sentiment periods while their predictive ability deteriorates when investor sentiment is low. It is important to note that both sentiment and attention measures can also be noisy. As Cookson et al. (2023) has argued, social media users behave the same way humans do in other settings: by following users who share their beliefs they build a personalized newsfeed that supports their original views.

Agents have initial views and then make choices about the information they collect. This is consistent with the study performed by Cookson et al. (2022), they conclude that although attention is highly correlated across different investment social media platforms, but sentiment is not. Cheon and Lee (2018) show that investors overpay for stocks with recent positive extreme returns due to the attention-grabbing feature of these stocks. Chen, He, Tao, and Yu (2023) find that mispricing due to "investor inattention" underlies many seemingly unrelated anomalies. They advocate for the incorporation of behavioral biases into an otherwise standard investment-based asset-pricing model to explain a broad set of anomalies.

3. Modeling Framework and Data

3.1. Unifying Econometric Modeling Framework

Our principal methodology is a general multivariate reduced-rank regression (RRR) model that encompasses principal component analysis (PCA), canonical correlation analysis (CCA), and vector autoregression (VAR)—tools that have been used in earlier studies—as special cases. Economic time series in general share some common characteristics such as trend, seasonality, and serial correlation. The existence of common elements is identified through the indicators of co-movement, latent or otherwise. The parsimonious structure that results in co-movement can be stated as in Hasbrouck and Seppi (2001) as,

$$\mathbf{r_t} = \mathbf{A}\mathbf{f_t} + \mathbf{a_t}, \tag{1}$$

where $\mathbf{r_t}$ is an m dimensional asset return vector, $\mathbf{f_t}$ is a lower dimensional ($r < m$) common feature vector, while the coefficient matrix $\mathbf{A}$ is of dimension $m \times r$. The $\mathbf{f_t}$ feature vector, if assumed latent, is constructed from the principal components of returns, or can be assumed known and related to exogenous variables $\mathbf{x_t}$, such as order flows. Thus, we can write

$$\mathbf{f_t} = \mathbf{B}\mathbf{x_t} + \epsilon_\mathbf{t}, \tag{2}$$

where $\mathbf{B}$ is a $r \times n$ (where $r < n$) matrix. Combining (1) and (2), we arrive at the unifying model

$$\mathbf{r_t} \quad = \mathbf{C x_t} + \mathbf{e_t} = \mathbf{AB x_t} + \mathbf{e_t}. \tag{3}$$

The model (3) exhibits multivariate features because of the rank constraint on the coefficient matrix, $\mathbf{C}$. The rank of the matrix $\mathbf{C}$ can be taken to indicate the effective number of common factors driving the returns and order flow relationships. The component matrix $\mathbf{B}$ can also be readily related to the principal components or the canonical correlation coefficients.

It can be shown that the matrices $\mathbf{A}$ and $\mathbf{B}$ can be obtained from the singular value decomposition of a standardized version of the matrix $\mathbf{C}$, or from the eigenvectors of

$$\begin{aligned} \mathbf{W} &= \mathbf{\Gamma^{1/2} \Sigma_{rx} \Sigma_{xx}^{-1} \Sigma_{xr} \Gamma^{1/2}} \\ &= \left(\mathbf{\Gamma^{1/2} C \Sigma_{xx}^{1/2}}\right) \left(\mathbf{\Sigma_{xx}^{1/2} C' \Gamma^{1/2}}\right), \end{aligned}$$

where $\mathbf{\Sigma}$'s are the covariance matrices. If $\mathbf{\Gamma = \Sigma_{rr}^{-1}}$, we obtain the results related to CCA, and if we set $\mathbf{x_t} = \mathbf{r_t}$ and choose the weight matrix to be $\mathbf{\Gamma = I_m}$, then we obtain the results for PCA. More precisely, if 'V_j' is the normalized eigenvector that corresponds to the j^{th} largest eigenvalue (squared canonical correlation) λ_j^2 of the matrix $\mathbf{W}$ above, then

$$\mathbf{A = \Gamma^{-1/2} V}, \qquad \mathbf{B = V' \Gamma^{1/2} \Sigma_{rx} \Sigma_{xx}^{-1}} \tag{4}$$

where $\mathbf{V} = [V_1, ..., V_r]$. We note in passing that the matrices $\mathbf{\bar{A} = V' \Gamma^{1/2}}$ and $\mathbf{B}$ above (when $\mathbf{\Gamma = \Sigma_{rr}^{-1}}$) are exactly the same canonical vectors given in Hasbrouck and Seppi (2001). When $\mathbf{\Gamma = I_m}$, and $\mathbf{x_t} = \mathbf{r_t}$, the eigenvalues and eigenvectors are based on the variance-covariance matrix of $\mathbf{r_t}$. In fact, the matrix $\mathbf{V}$ contains the first 'r' principal components, and $\mathbf{B = V'}$ in (3).

A complementary aspect of the reduced-rank formulation of (3) is that there are $(m - r)$ linear constraints on the coefficient matrix $l_i' \mathbf{C} = 0$. This implies that $l_i' \mathbf{r_t} \sim l_i' \mathbf{e_t}$, the linear combinations of returns that are independent of $\mathbf{x_t}$, where l is essentially the eigenvector that corresponds to the smallest canonical correlation between the $\mathbf{r_t}$ and $\mathbf{x_t}$. These relationships

among $\mathbf{r_t}$ are sometimes referred to as structural relationships that will remain stable, and do not depend on $\mathbf{x_t}$ over time. The multivariate regression formulation in (3) is flexible and can include other variables that might potentially affect commonality. As an example, it could include investors' sentiment or attention scores, an aspect that we will explore in later sections as the potential sources of commonality. Observe that if $\mathbf{x_t} = \mathbf{r_{t-1}}$, model (3) is essentially VAR(1) with constraints on stationarity in addition to possibly rank constraints.

Note that the general model specified in equation (3) has recently found quite a few applications in finance. For instance, the formulation of Black CAPM leads naturally to a reduced-rank regression model with coefficient matrix of unit rank (see Velu and Zhou (1999)). More recently He, Huang, Li, and Zhou (2022) propose a reduced-rank approach to reduce a large number of factors to a few parsimonious ones that can better explain the cross-sectional variation of stock returns. Wan, Li, Lu, and Song (2023) demonstrate that multivariate analysis using reduced-rank regression can make model estimation more robust, flexible, and statistically accurate. Adrian, Crump, and Vogt (2019) study the non-linear risk-return trade-off by relating stock market volatility to future returns, using a nonparametric extension of the model (3) called sieve reduced-rank regression. Kelly, Malamud, and Pedersen (2023) provide an asset-pricing framework based on the eigenvectors of a prediction matrix to construct what they call "principal portfolios", which produces significant out-of-sample alphas. The portfolio construction methodology also involves concepts related to reduced-rank regression. For other applications of the reduced-rank regression model, see Reinsel, Velu, and Chen (2022).

3.2. Data Summary

Return and Order Flow

The return and order flow data are obtained from the NYSE's Daily Trade and Quote (TAQ) database. We focus our analysis on a fairly large sample of 1,312 stocks with large market capitalization in the US. We apply our study to cover an extended time period of nine calendar years, as one of our primary goals is to investigate the long duration persistence of commonality. Our sample consists of stocks that are liquidly traded predominantly by institutional traders,

and should therefore exhibit more commonality.

We aggregate the intraday tick-level data from TAQ into 15-minute intervals, resulting in 26 observations per trading day from 9:30am to 4:00pm. The chosen time resolution of 15-minute interval can smooth out high-frequency fluctuations that result from pure noise trading and momentary market imbalance. It is also easier to model the relationship between returns and order flows in discrete time intervals. Returns are calculated as the difference in the log midpoint quotes over the two endpoints of each interval as $r_t = \log M_t - \log M_{t-1}$, where $M_t = \frac{\text{ask}_t + \text{bid}_t}{2}$. For order flows, we compute both signed and unsigned measures. For ease of notation, we drop the firm index 'i'. For the j^{th} trade within each interval, let P_j and v_j denote the dollar price per share and share volume. Four unsigned order flow measures (number of trades, share volume, dollar volume, and square root of dollar volume) are derived from the consolidated trade data for each time interval. Defining $\text{sign}(v_j) = 1$ for a "buy" trade and $\text{sign}(v_j) = -1$ for a "sell" trade, we can also construct signed order flow measures. We use the matching algorithm by Lee and Ready (1991) to classify the direction of each trade. Let n_t denote the total number of trades within an interval, we follow Hasbrouck and Seppi (2001) to compute the following signed and unsigned order flow measures:

- Trades: n_t and $\sum_{j=1}^{n_t} \text{sign}(v_j)$
- Share volume: $\sum_{j=1}^{n_t} v_j$ and $\sum_{j=1}^{n_t} \text{sign}(v_j) \cdot v_j$
- Dollar volume: $\sum_{j=1}^{n_t} \log P_j \cdot v_j$ and $\sum_{j=1}^{n_t} \text{sign}(v_j) \cdot \log P_j \cdot v_j$
- Square root dollar volume: $\sum_{j=1}^{n_t} \sqrt{\log P_j \cdot v_j}$ and $\sum_{j=1}^{n_t} \text{sign}(v_j) \cdot \sqrt{\log P_j \cdot v_j}$

Investors' Sentiment and Attention

Financial news can be broadly classified into two main groups: "structured" and "unstructured". Specific financial news such as earning statements are released on a regular basis, and models to analyze such events are well-established in the literature. Research focus has recently shifted to the analysis of many other unstructured information, such as news related to the products or personnel of companies that may arrive unscheduled, and that carry important

information for trading decisions. These news-based event strategies are fairly well-studied in the finance literature (see, for instance, Calomiris and Mamaysky (2019) and Jiang et al. (2022)).

A salient property of unstructured news streams is that they come at irregular intervals, are usually qualitative in nature, and are in a raw text format. In order to make use of these unstructured news as an information source to link to stock performance, the signals from the textual data need to be appropriately quantified. Because the news often originate from multiple sources, including noisy social media, they need to be properly aggregated in order to extract the signal from the noisy data.

Das and Chen (2007) show how it is possible to capture the net sentiment from positive and negative views on message boards using statistical natural language processing (NLP) techniques. For instance, by relating the sentiment to the performance of stocks in the Morgan-Stanley High-Tech Index, they report that although there is no dominant relationship to individual stock prices, there is a statistical relation to the aggregate index performance. There is also a strong relationship between message volume and return volatility, as reported by Antweiler and Frank (2004). These examples clearly illustrate that investors sentiments expressed via the social media and other platforms can exert a noticeable impact on stock market behavior.

A number of web analytics companies[1] have sprung up in the last decade, offering alternative data analytic services by aggregating the web information and processing sentiment data related to individual stocks. The data from SMA[2] is made available to us for the period we used in our study (2012 through to 2020). The data based on Twitter feed contains both raw and processed stock level sentiment and attention measures at a 15-minute interval. The data is aggregated to introduce a 5-minute time offset with respect to the TAQ data on return and order flow measures, so that the daily 26 observations span from 9:25am to 3:55pm. Recall that the first TAQ data interval is based on trades executed within the 15-minute interval between 9:30am and 9:45am. On the other hand, the first SMA data interval is based on Twitter information within a rolling 24-hour window up until 9:25am. This 5-minute time offset allows us to assess whether this unstructured information has predictive power on return and order flow measures.

[1] iSentium, SMA, MarketPsych, RavenPack, to name a few.

[2] The authors of this paper have no affiliation with SMA.

The SMA database contains useful proxies for investors' sentiment and attention measures. In our study, we use the "raw sentiment Z-score" as the sentiment proxy. This is computed first by summing up the sentiment of unique tweets from credible Twitter accounts in a rolling 24-hour window. This raw sentiment score is then transformed by using a 20-day moving average of mean and volatility measures to generate the statistical Z-score of the raw sentiment score. For attention proxy[3], we use "tweet volume", defined as the number of unique tweets arriving within a 24-hour interval.

Note that the sentiment proxy is by definition a "signed" measure—the raw sentiment Z-score can be either positive or negative. Their sign and magnitude provide an indication about the sentiment that investors attribute to a given stock. Conversely, the attention proxies are by definition "unsigned" measures— tweet volume and buzz can only vary on the positive range. By intuition, one should expect that extreme sentiments (very positive or very negative) ought to be accompanied by very high attention, while mild sentiments occurs when attention is moderate. The analyses will provide empirical evidence for this nonlinear relationship between sentiment and attention measures. Our results also reveal the linear (nonlinear) associated between returns and order flows commonality *vs* attention (sentiment) measures, demonstrating the multi-dimensional aspect of commonality relationship.

3.3. Descriptive Statistics

As mentioned earlier, the data on returns and order flows are obtained from TAQ, while our sentiment and attention data are from SMA. Between the two sources, we have a total of 1,342 overlapping stocks, and among them 1,213 are listed on NYSE, AMEX and NASDAQ[4]. We focus our analysis on these 1,213 stocks, which account for, on average, 68.8% of the total market capitalization over the sample period from Jan-2012 to Dec-2020.

In Table 1 the descriptive statistics of the measures are summarized over the nine calendar years. We compute returns from the log midpoint quotes, and use both signed and unsigned number of trades to measure order flow. The sentiment proxy is based on raw sentiment Z-

[3]We also tested using "buzz" as our attention proxy, defined as a measure of unusual volume activity compared to a universe of stocks, and obtained similar empirical results.

[4]Stocks with Exchange codes 1, 2, and 3 in CRSP.

score, and the attention proxy from tweet volume. As a comparison, we also present descriptive statistics for S&P 500 and VIX indexes in the table. Overall, the sample stocks exhibit higher average return compared to the S&P 500 index with higher volatility. Note also that the number of trades has increased over the decade, which can be attributed to increasing prevalence of algorithmic trading. Tweet volume varies over time, but in general exhibits an upward trend. This is likely due to the growth of Twitter as a communication platform by investment analysts over recent years.

In order to provide a graphical illustration of the variation in the measures used, Figure 1 plots select variables for Jan-2020 to Jun-2020, a period that includes the 2020 stock market crash, when the financial market experienced a spike in volatility, that was followed by a quick recovery. The top panel shows the dip (spike) and recovery (fall) in the S&P 500 and VIX index levels, while the middle panel shows the daily return and turnover of the 1,213 stocks. The bottom panel charts the average sentiment and attention measures—highlighting the overall fall and recovery in investors' sentiment, along with a corresponding rise and fall in their attention. The daily sentiment score decreases with S&P 500 index in March 2020, and recovers subsequently in April 2020. The daily total tweet volume increases 33% from 6,000 to 8,000 in January and February, maintains above 8,000 in March and April, before falling back to below the 8,000 level in May and June. During this period, the selected stocks average return is highly correlated with the S&P 500 index return (0.93) and with the average sentiment score (0.43). The VIX index is negatively correlated with the average sentiment score (-0.38) as expected, but positively correlated with tweet volume (0.24).

Our analysis reveals a high degree of correlation among the order flow measures. This is an indication that these measures all capture closely related information. Given this observation, we could in theory use the first principal component score as an index measure of order flows on the right hand side of the market impact model in equation (3). Recall that market impact models measure the impact of trading activities on price movement, and is usually measured by the regression of price change on the ratio of trade size to market size. However, using an index measure (such as a principal component) might render our results difficult to interpret, and hence we instead look for the single most representative order flow measure that exhibit

the highest correlation with the return.

Most market microstructure research have identified a positive association between a "buy" ("sell") order with a price rise (fall). In this case, it is likely that signed order flow measures will capture information that is already partially captured by the return measure, while unsigned order flow measures can capture the full trade volume information that is not accounted for by return. In fact, signed trades are generally known to be the most highly correlated with returns at the individual stock level[5]. Given that the order flow measures are highly correlated , choosing any one of these order flow measure will suffice. In subsequent analysis, we will therefore be focusing on unsigned trades as the order flow measure.

4. Commonality in returns, order flows, sentiment, and attention

4.1. Pre-Modeling Analyses

Prior to presenting results of the unifying reduced-rank regression model, we begin with other methods that are common in the literature. This will help us to connect with the past studies.

Principal Component Analysis

We run principal component analysis on our full data sample to infer the evidence for the existence of commonality in our measures. The variables are standardized, thus the PCA is performed on the correlation matrix, so its eigenvalues add up to the number of stocks. If there is a high correlation among the variables, the proportion of the variance explained by the first new principal components should dominate. On the other hand, if the variables are uncorrelated, then we can expect each component to have approximately equal contribution—in other words, each principal component explains roughly 0.082% of the total variance (e.g. $1/1,213 \approx 0.082\%$). This rule of thumb provides a way to interpret the magnitude of the eigenvalues that result from the PCA analysis. In financial data, the variance explained by the first

[5]Hasbrouck and Seppi (2001) explore both the square root dollar volume and the trade size as representative measures, while Harford and Kaul (2005) use signed volume or signed number of trades in their study.

few principal components is typically large. Hence, following the convention of the literature, we report in Table 2 only the variance explained by first three principal components of return, order flow, sentiment, and attention measures. The dominance of the three components support the hypothesis on commonality among the measures used in the dataset. These measures have also been adjusted to remove the time-of-day effects documented in Wood, McInish, and Ord (1985). With 1,213 sample stocks under consideration, the cumulative variance explained are markedly higher than the mean level of $\frac{1}{1213} \approx 0.082\%$.

To investigate the persistence of the principal components, we study their behavior over time. Figure 2 shows the principal component analysis on a 12-month moving window for return (top) and both signed (middle) and unsigned trade (bottom) measures. In the literature, the variance explained by the dominant principal components is often taken as an indicative measure on commonality. As Longin and Solnik (2002) have demonstrated, correlations in the equity market is related to the market trend—they increase during bear markets, but decrease during bull markets. Overall, there is a longitudinal upward trend of in commonality. As one would expect, the return commonality index experiences a significant upward jump during early 2020 due to the significant Covid-19 pandemic related market crash. This jump is also reflected in the signed and unsigned trade measures.

Canonical Correlation Analysis

The existence of commonality among assets manifesting in return and order flow measures individually are conclusively established by the preceding analysis. In this section, we use Canonical Correlation Analysis (CCA) to investigate the relationship between return and order flow measures. If the two sets of variables are statistically related, a direct way to study this correlation is through CCA, as expounded in Hasbrouck and Seppi (2001).

In fact, building on insights from the single-asset return-order flow relationship established in earlier market microstructure research, Hasbrouck and Seppi (2001) postulate that the commonality in order flows is the source of commonality in returns. A large number of subsequent research in the commonality literature have also identified this strong relationship. The our

longitudinal analysis below will reveal that the largest canonical correlation between returns and order flows is consistently high over the years with minimal variation over time. This observation suggests that these two variables could be moving together instead and are determined endogenously, while their relationship may be explained by other exogenous factors.

The first and third canonical correlations for our sample period are plotted in Figure 3. The first canonical correlation can vary from 0.94 to 0.97 with no discernible patterns on any years, even during pandemic period when the commonality index based on returns was higher than other periods. Although there is moderately more variation in the third canonical correlation over time, the fluctuation cannot be readily associated with any easily identifiable economic events, and the variations are also mild, suggesting a stable relationship between returns and order flows over the decade. The high canonical correlations indicate that returns and order flows have strong co-movement, and are not likely to drive each other.

The canonical correlation coefficients presented in Table 3 confirm that there are extensive contemporaneous effects among the variables. Here "Ret" is the log midpoint returns, "Trade" is the signed trades measure, "Sent" is the tweet sentiment Z-score (sentiment measure), and "TVolume" is the tweet volume (attention measure). For comparison, we have also added a row for "TBuzz", which measures the amount of unusual volume activity, which can also be interpreted as an alternative measure for investors' attention as our results suggest. Apart from the CCA between return and order flow measures, we also compute their CCA with respect to sentiment and attention measures. In order to investigate the direction of information flow, we also compute CCA coefficients for contemporaneous, lag-1, and lag-12 observations. The lead-lag canonical correlations between return, order flow, attention, and sentiment measures shed light on the direction of information flow between pairs of variables. Note that a high CCA coefficient for contemporaneous observations merely suggest a high degree of contemporaneous correlation. However, if a high CCA coefficient exists between a variable with another lagged variable, then the lagged variable can be interpreted as exhibiting predictive power. Our CCA analyses reveal that although contemporaneous return and order flow exhibit a high CCA coefficient of 0.879, this relationship falls steeply to 0.584 for the CCA between return and lag-1 order flow. On the other hand, the CCA between return *vs* sentiment and return *vs* attention

measures remain high in the range of 0.85, even after introducing multi-period lags. Similarly, the CCA between order flow *vs* sentiment and attention measures remain strong even when we introduces lagged variables. Therefore we can conclude that our results strongly suggest that return and order flow are contemporaneous measures of similar information—one does not predict the other. However, sentiment and attention measures both exhibit predictive power over return and order flow. In other words, the lagged sentiment and attention measures can impact current return and order flow. This observation is noteworthy, considering the way SMA dataset is aggregated; it has introduced a 5-minute lag in the attention and sentiment data with respect to the return and order flow dataset.

Next we extend our analysis to include both investors' attention and sentiment measures, and explore their relationship with returns and order flows. Figure 4 presents the time series of standardized sentiment score. To remove time-of-day effect, we normalize the sentiment score for each stock (1,213 stocks) and each time-interval (26 time intervals) for the whole sample. For the results with 12-month rolling window, we collect the raw data (12 months) and normalize the collected data rather than use the standardized data generated from full-sample. This figure provides a clear longitudinal illustration on the cross-sectional variation of the sentiment measures over the nine calendar years we have studied. Two important observations immediately stand out in this figure. First, the distribution on the range of sentiments is unevenly split between positive and negative—on aggregate, there are significantly more positive sentiments than negative ones. Second, the distribution is right skewed, displaying a heavy tail on positive sentiments, and a light tail on negative sentiments. Nevertheless, we note that both observations are intuitive and comprehensible. Financial analysts are expected to be more vocal in positive news and sentiments, and at the same time exhibit beyond normal exuberance and can even exaggerate their messages. On the other hand, when tweeting negative news or sentiments on social media, they are expected in general to be more conservative and adopt a more reserved tone.

Having introduced investors' attention and sentiment measures as exogenous factors to our analysis, we proceed to study their relationships with respect to returns and order flows. First, we employ canonical correlation analyses to capture their contemporaneous and lagged relation-

ships, for which Figure 5 provides a graphical illustration. The canonical correlation coefficients between contemporaneous returns and order flows are high and comparable to the canonical correlations between returns and sentiment or attention. The canonical correlations between order flows and sentiment or attention are also relatively stable over time. We argue that the persistently high canonical correlations observed for most recent nine years might also indicate that there may be exogenous economic or behavioral (sentiment or attention) factors driving both returns and order flows simultaneously. Indeed, although the CCA between contemporaneous returns and order flows is on average 0.95, between returns and order flows with lag-1 dropped to approximately 0.77. On the other hand, the CCA between returns and sentiment (as well as between order flows and attention) virtually does not vary when after introducing a lag. This is an important empirical result supporting our hypothesis that sentiment and attention measures are the exogenous factors exhibiting predictive power over the endogenous returns and order flows.

As discussed in the previous section, one can measure commonality through different methodologies using either principal component analysis, canonical correlation analysis, or reduced-rank regression. The distinction between these methodologies can be observed as follows: while PCA measures commonality among a set of related variables, the other two methods focus on joint or relational commonality of two or more sets of variables. We define reduced-rank return commonality as $\bar{\mathbf{A}}\mathbf{r_t} = \mathbf{V_1'}\boldsymbol{\Sigma}_{\epsilon\epsilon}^{-\mathbf{1/2}}\mathbf{r_t}$ where $\mathbf{V_1}$ is the eigenvectors associated with the largest eigenvalue of

$$\mathbf{W} = \boldsymbol{\Sigma}_{\epsilon\epsilon}^{-\mathbf{1/2}}\boldsymbol{\Sigma}_{\mathbf{rx}}\boldsymbol{\Sigma}_{\mathbf{xx}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\mathbf{xr}}\boldsymbol{\Sigma}_{\epsilon\epsilon}^{-\mathbf{1/2}},$$

where $\boldsymbol{\Sigma}_{\epsilon\epsilon}$ is the residual covariance matrix from the OLS regression $\mathbf{r_t}$ on $\mathbf{x_t}$. Empirically, our reduced-rank return commonality as measured by $\bar{\mathbf{A}}\mathbf{r_t}$ is highly correlated with the first PC score in return (with a correlation coefficient of 0.9442), but the correlation between reduced-rank loadings $\bar{\mathbf{A}}$ and the first PC coefficients is 0.0749. Similarly, the correlation between reduced-rank order flow commonality $\mathbf{Bx_t}$ and the first PC score in order flow is 0.8567, and the loadings correlation is 0.0805. These indicate that the indices constructed out of the two method may be composed of different weights, but overall they measure similar concepts and

effects. Note that our unifying reduced-rank regression model in Equation (3) can also capture the dimension of this relationship—the variation in returns $\mathbf{r_t}$ can be modeled through the variation in order flows $\mathbf{x_t}$, the rationale being that asset-related information is likely to be absorbed in order flows first, and this in turn will lead to the variations in returns.

Co-integrated Regression

Finally, we perform a regression version of co-integration analysis to lend further support to our hypothesis. We note that it is a common practice to study co-movement for stationary time series and co-integration for non-stationary time series. If a longitudinal variable vector is co-integrated, which implies that there are some linear combinations of the vector that are weakly dependent, then these variables can be taken as somewhat redundant, and any deviation of their value from the linear combination of the other variables is temporary and can be expected to revert. Although the concept of co-integration generally applies to non-stationary vector series, it can be conceptually extended to the regression model in Equation (3) as follows. If we write the model (3) in the so called error-correction form

$$\mathbf{r_t} - \mathbf{x_t} = (\mathbf{C} - \mathbf{I})\mathbf{x_t} + \epsilon_\mathbf{t}, \tag{5}$$

we can use the smallest canonical correlation between $\mathbf{r_t} - \mathbf{x_t}$ and $\mathbf{x_t}$ to construct linear combination of $\mathbf{r_t} - \mathbf{x_t}$ (following the relationship $l_i'\mathbf{C} = 0$) that is independent of $\mathbf{x_t}$, so that $l'\mathbf{r_t} \sim l'\mathbf{x_t}$.

Figure 6 is plotted based on this formulation, and clearly indicates that this is a viable relationship. The R^2 of the univariate regression (red line) is 0.5732. This is yet another evidence supporting our argument that both returns and order flows (indirectly volume) have common sources of influence, as has been pointed out in the seminal work of Tauchen and Pitts (1983), where volume and volatility are shown to be related in the presence of varying information.

4.2. Reduced-Rank Regression Analysis

The discussion earlier related to the commonality index provides some insights into the relationship between the performance of individual stock and the performance of the market overall. However, it is important to seek an economic interpretation, of the coefficients of the principal components, or the component matrices $\mathbf{A}$ and $\mathbf{B}$. After all, the matrices $\mathbf{A}$ and $\mathbf{B}$ are directly based on the dynamic relationship between returns and order flows. As our results have demonstrated, the first PC's loadings of returns are roughly equal and thus, they reflect the composition of the general market factor. By simply examining these coefficients, we do not see that larger firms get more weights consistently. As the constituent stocks might vary over time in the Dow Jones index, we check on the consistency over time via the cosine measure proposed by Krzanowski (1979)[6]. Using this cosine measure and taking the average correlation over the years, we obtain a similarity index for the returns. The values range above 0.93 and clearly demonstrate that there is a remarkable similarity of the coefficients over time. The consistency generally holds for the comparison between $\mathbf{A}$ and $\mathbf{B}$ estimates over time as well.

In order to see how our reduced-rank regression methodology unifies previous modeling framework, note the resemblance between the principal components of $\mathbf{r_t}$ and the model coefficients in (3). Observe that if $\bar{\mathbf{A}}$ is a reflexive generalized inverse[7] of $\mathbf{A}$, then we can rewrite Model (3) in the following form:

$$\bar{\mathbf{A}}\mathbf{r_t} = \mathbf{B}\mathbf{x_t} + \bar{\mathbf{A}}\mathbf{e_t}. \tag{6}$$

If there is no relationship between $\mathbf{r_t}$ and $\mathbf{x_t}$ (i.e., $\mathbf{B} = \mathbf{0}$), then the principal components would be proportional to the rows of $\bar{\mathbf{A}}$ matrix. The difference between the principal components and the $\bar{\mathbf{A}}$ matrix is that $\bar{\mathbf{A}}$ is resulting from the joint relationship between $\mathbf{r_t}$ and $\mathbf{x_t}$.

It is also important to note that there is a difference between principal components resulting from model (1) and the canonical components of the returns ($\bar{\mathbf{A}}$) in model (6). The former are obtained not from the joint relationship with order flows and therefore the resemblance is not strong. This further confirms the fact that simply relating principal components based on

[6] If $\mathbf{L}$ and $\mathbf{M}$ are $m \times k$ matrices of the principal component loadings from two years, say, then the similarity between the two is given by $\cos^{-1}(\sqrt{\lambda_i})$ where 'λ_i' is the largest eigenvalue of $\mathbf{S} = \mathbf{L}'\mathbf{M}\mathbf{M}'\mathbf{L}$. If $k = 1$, λ_1 is simply the square of inner product of the two vectors.

[7] $\mathbf{A}\bar{\mathbf{A}}\mathbf{A} = \mathbf{A}$ and $\bar{\mathbf{A}}\mathbf{A}\bar{\mathbf{A}} = \bar{\mathbf{A}}$

individual set of variable may not be optimal if the focus is on the joint relationship among two or more sets of variables. Thus we advocate using the reduced-rank regression methodology directly instead of these process of relating the principal components of returns to principal components of order flows.

Exploring the Index Constituents Hypothesis

We explore the reduced-rank loadings $\bar{\mathbf{A}} = \mathbf{V}_1' \boldsymbol{\Sigma}_{\epsilon\epsilon}^{-1/2}$ for S&P 500 constituents and non S&P 500 constituents. Here, we rescale the vector so that $\bar{\mathbf{A}}'\mathbf{1} = \mathbf{1}$—this allows us to investigate the weight on individual stocks to form a market portfolio obtained from the reduced-rank model. If there is no difference in commonality behavior between an S&P 500 constituent stock and a non-S&P 500 constituent stock, then we should observe no material difference in a stock's weight whether it is in the S&P500 index or not. As mentioned in the literature review, it has been established that inclusion of a stock into a major equity index can lead to an increase in its co-movement with the other constituent stocks (see, for instance, Green and Hwang (2009)). This can be attributed to: (1) trading stemming from passive index-tracking funds will generate order flows on stocks included into the index as they mimic the index portfolio, and (2) market participants' responses to sentiment and attention related to the S&P index as a whole will drive the interest in stocks included in the index.

Under our reduced-rank model, the average weight for S&P 500 constituent and non-constituent stocks is positive and negative, respectively. This is attributed to the fact that there is a stronger returns (and order flows) commonality among S&P 500 constituent stocks due to the stronger liquidity demand and trading activity, while commonality for the other non-constituent stocks are more diversified. This provides us a way to study the change in a stock's commonality when it is included to (or excluded from) the S&P500 index constituent by tracking its weight in the standardized $\bar{\mathbf{A}}$.

In our data set, we have identified 59 stocks that have been added to the S&P 500 index and 39 stocks removed from the S&P index during the time period of our study. Using the reduced-rank formulation above to compute the weights on these stocks 98 stocks in the periods before

and after the inclusion, we are able to study the effect of index inclusion/exclusion on the changes of commonality. Figure 7 displays how the weight of added/removed stock changes before and after the event. On aggregate, the weight of the added stock begins to increase 5 months before it is added and reaches the peak in 4 months after it is added. The change in weight is statistically significant, with a p-value of 0.0515. For a stock that is removed, the weight starts to decrease from 6 months before it is removed and hit the trough in 7 months after it is removed, and again the weight difference is statistically significant with a p-value of 0.042.

Comments on Commonality and Co-movement

In the study of joint behavior of multiple stocks, there is an overlapping discussion about commonality and co-movement. For example, in the studies of price-based co-movement, it is expected that higher priced stocks exhibit a certain degree of similarity, and there is some commonality among their behavior. The co-movement is modeled via VAR models in the form of

$$\mathbf{r_t} = \mathbf{\Phi} \mathbf{r_{t-1}} + \mathbf{e_t}, \tag{7}$$

i.e. a regression model of type (3), but with lag-1 returns $\mathbf{r_{t-1}}$ in place of $\mathbf{x_t}$. An appropriate linear combination of $\mathbf{r_t}$ will be free of this co-movement when the dependency is accounted for. This is typically done via a reduced-rank VAR model, where the coefficient matrix, $\mathbf{\Phi}$, is of reduced rank. From the relationship $l_i'\mathbf{C} = 0$, it can be seen that $l'\mathbf{\Phi} = 0$ produces the linear combination $l'\mathbf{r_t}$ that is free of the serial dependence. In other words, co-movement is a stronger restriction than commonality as discussed in the literature.

The VAR(1) model on the returns has been considered by several authors. Among them, DeMiguel, Nogales, and Uppal (2014) study the serial dependence of returns and show empirically that the mean-variance portfolios based on the predicted values of returns from VAR model outperform traditional portfolios formed from past return characteristics. In this work, we want to explore whether the model in equation (3) can be improved by accounting for the

serial correlation in the residuals. This can be incorporated via the model below:

$$\mathbf{r_t} = \mathbf{C}\mathbf{x_t} + \mathbf{D}\mathbf{r_{t-1}} + \mathbf{u_t} \tag{8}$$

Our analysis reveals that there is not much difference between the two models (7) and (8). This is potentially due to the fact that co-movement refers to time series behavior, while commonality refers to cross-sectional behavior. The off-diagonal elements of the $\mathbf{D}$ matrix that indicate dependence across stocks are generally not significant, where only 11.02% off-diagonal elements among the 1,213 stocks are significant at 0.05 level.

5. Sources of commonality

Existing literature has identified that the dominant driver of return and order flow commonality are 1) information diffusion, and 2) common holdings by investors. Even though it is not possible to know the exact composition of every individual investor's portfolio, their trading activities over a duration may indicate how commonality can arise. Because trading is driven primarily by information (including noise) via the influence of sentiment and attention, we focus our analysis on the information diffusion aspect. With the advent of internet and social media, information (albeit noisy one) gets disseminated almost immediately. In our study, we focus on the use of twitter data in extracting stock-related information via the appropriate natural language processing methods. The analysis of investor sentiment data to provide market signals has been a focus also in recent research on trading.

We assimilate existing views and insights on the importance of investors' sentiment and attention measures into a more general framework. Instead of considering them as alternative metrics, we argue that the two measures capture different aspect of investors' behavioral characteristics, and they exhibit a nonlinear mutual relationship. Empirically, high attention is associated with strong (positive or negative) sentiment, while low attention is associated with weak (muted or neutral) sentiment. Based on this observation, we postulate the hypothesis that the noise in sentiment measure is higher (lower) for stocks under high (low) attention—this is due to confirmatory bias, where herding mentality results in investors echoing sentiments similar to

their own for stocks under high attention. On the other hand, the noise in the attention measure should be higher (lower) for stocks with neutral (strongly positive or negative) sentiment—strong sentiments are attention grabbing, while neutral sentiment leads to inattention instead. Both measures are capable of providing useful information on return and order flow.

A main objective of our paper is to provide further insight to the commonality between return and order flow that is widely reported in the literature. We first demonstrate that return and order flow are simply contemporaneous measures—their relationship weakens once a lead-lag period is introduced. We therefore model both return and order flow as endogenous variables, and argue that they are driven by the exogenous investor's sentiment and attention. We elaborate on how sentiment and attention data collected and processed in real time can help to explain the commonality in both returns and order flows in a high frequency setting.

We test the hypothesis that return is associated with sentiment, while order flow is associated with attention. The relationship between return and order flow is driven by the association between sentiment and attention measures. Investors buy or sell a security based on the sentiment they held on the stock, which in turn leads to a positive or negative return. On the other hand, trading volume is driven primarily by investors' attention—stocks under high attention experience large trading volume, but inattention leads to low trading activity. Our postulate is also based on empirical observation on the relationship between return and trading volume reported in Han et al. (2022). Our results provide a strong empirical confirmation that the persistence nature in the way investors interpret market sentiment and attention as an important source of commonality. A major advantage of our proposed approach from gauging sentiment and attention based on news analytics is that this relationship can be exploited for the development of real-time trading strategies as well.

To briefly illustrate the persistence of commonality, we focus on two of the sources identified in the literature and stated in the first part of Section 2. These are stocks that belong to an index or stocks that are grouped into price tiers. To add figures.

In this section, we lay out and test our hypothesis that investors' sentiment and attention consistently constitute as the sources of commonality. We test our hypothesis using the following

multivariate regression model:

$$\begin{bmatrix} \mathbf{r_t} \\ \mathbf{x_t} \end{bmatrix} = \mathbf{C} \begin{bmatrix} \mathbf{s_t} \\ \mathbf{a_t} \end{bmatrix} + \mathbf{u_t} = \mathbf{AB} \begin{bmatrix} \mathbf{s_t} \\ \mathbf{a_t} \end{bmatrix} + \mathbf{u_t} \tag{9}$$

where $\mathbf{s_t}$ is the stock-level sentiment score and $\mathbf{a_t}$ is the stock-level tweet volume (attention) score. Both have been adjusted to remove the intraday effect. Note with $m = 1,213$ stocks, the regression coefficient matrix, $\mathbf{C}$, is of large size $2m \times 2m$. Because of commonality in the relationships, we should expect the rank of the $\mathbf{C}$ matrix to be much smaller than $2m$. To give an indication of the rank of C, or the dimension of $\bar{\mathbf{A}}$. We construct the return and order flow commonality as $\mathbf{r_t}^{\text{com}} = \bar{\mathbf{A}}_\mathbf{1}\mathbf{r_t}$ and $\mathbf{x_t}^{\text{com}} = \bar{\mathbf{A}}_\mathbf{2}\mathbf{x_t}$ where $\bar{\mathbf{A}} = \left[\bar{\mathbf{A}}_\mathbf{1}, \bar{\mathbf{A}}_\mathbf{2}\right]$ is a reflexive generalized inverse of $\mathbf{A}$. We also create the aggregate sentiment index and attention index as $\mathbf{s_t}^{\text{agg}} = \mathbf{B_1 s_t}$, $\mathbf{a_t}^{\text{agg}} = \mathbf{B_2 a_t}$ where $\mathbf{B} = [\mathbf{B_1}, \mathbf{B_2}]$. Figure 8 presents the time series plots for the high-frequency sentiment and attention indexes. Sentiment index exhibits volatility over time, but experience more volatile swings in recent period. Attention index is increasing in trend, reaching its peak during the 2020 pandemic period. The correlation between our sentiment index and attention index is -0.256, an indication that they probably capture different dimensions of information. We fit a GARCH(1, 1) model for the indices. Both the sentiment and attention time series are non-stationary:

Sentiment Index:

$$s_t = \epsilon_{S,t} \quad \text{where } \epsilon_{S,t} = \sigma_{S,t} z_t^s$$
$$\sigma^2_{S,t} = 0.05 + 0.02\sigma^2_{S,t-1} + 0.92\epsilon^2_{S,t-1}$$

Attention Index:

$$a_t = \epsilon_{A,t} \quad \text{where } \epsilon_{A,t} = \sigma_{A,t} z_t^a$$
$$\sigma^2_{A,t} = 0.005 + 0.99\epsilon^2_{A,t-1}$$

Fit GARCH(1, 1) here to present the results. Put an asterisk on the significant coefficients.

Figure 9 provides a graphical illustration on the relationship between the attention and sentiment indexes based on the reduced-rank regression framework. The top figure shows the

scatter plot of the attention index *vs* the sentiment index, with the LOWESS plot overlaid in red. The bottom figure is an aggregate of the sentiment index into quintiles, and plot attention *vs* sentiment against the average of the corresponding attention index. These plots clearly demonstrate the nonlinear relationship between the two indexes. Intuitively, when investors' attention is high, the stocks exhibit strong sentiment—this can be either very negative sentiment (the 1^{st} quintile), or very positive sentiment (the 5^{th} quintile). In other words, there is very few "neutral" sentiment when attention is high and it appears to be polarized. On the other hand, when the sentiment index is moderate (from the 2^{nd} to the 4^{th} quintile), the attention index is correspondingly low. Next, we present an alternative visualization of our results to shed more light on this nonlinear relationship. Figure 9 also shows the scatter plot of sentiment index *vs* the attention index (i.e. Figure 9 rotated by $90°$), with the LOWESS plot overlaid in red. The bottom figure here aggregates the attention index into quintiles, and plot them against the standard deviation of the corresponding sentiment index. This figure reveals yet another insightful aspect of the relationship between the two indexes—the sentiment index is visually less varied when investors' attention is low. On the other hand, stocks under high attention exhibit significantly higher variation in their sentiment index. How do we bring in this non-linearity into the modeling? Does non-linearity in the predictors matters? This discussion should move to descriptive part in Section 4.

This nonlinear relationship is insightful when we use investors' attention and sentiment as exogenous measures to drive the return and order flow commonality observed in the research literature. Academic research in market microstructure and behavioral finance have identified a relationship between investors' sentiment and the subsequent returns of a given stock. Another related but separate strand of recent research has identified that investors' attention, as a scare resource, captures a different dimension of market information (see Chen et al. (2022), Huang et al. (2019), and Jiang et al. (2022)). In this work, we take both sentiment and attention measures as exogenous variables in our modeling framework. In the preceding sections, we have demonstrated that returns and order flows are contemporaneous measures of the same information, and that lagged order flows does not exhibit predictive power over subsequent returns. We have also illustrated that both investors sentiment and attention exhibit predictive

power at the intraday trading frequency level. In this section we build on these findings to further explore this relationship.

Our multivariate reduced-rank regression model in Equation (9) allows us to go beyond individual stock-level analysis to investigate the cross-sectional relationship instead. The left column of Figure 10 shows the relationships between return and order flow commonality with respect to the sentiment index. The figure shows the scatter plot of return (top) and order flow (middle) commonality *vs* sentiment index. The lower figure aggregates the sentiment index into quintiles, and plot them against the mean of the corresponding return and order flow commonality *vs* sentiment index. These figures provide empirical evidence demonstrating a strong nonlinear relationship between investors' sentiment *vis-à-vis* return and order flow commonality. Return and order flow commonality are positive when the sentiment index is very negative (1^{st} quintile) or very positive (5^{th} quintile), but are negative when the sentiment index is moderate (2^{nd} to 4^{th} quintiles).

It can also be observed from the scatter plots that the dispersion in return and order flow commonality are high when sentiments are strong (1^{st} and 5^{th} quintiles), but are lower when sentiments are mild (2^{nd} to 4^{th} quintiles). This also suggests that market behaves in a highly concentrated way when investors' sentiment is strong, but becomes more diversified when sentiment is mild. This observation is consistent with the expectation that when trading activities, and hence order flow commonality, is highly correlated, it could either be caused by the investors' sentiment on the stock market being very positive or negative on the whole. On the other hand, when order flow commonality is low, investors' sentiment is more concentrated around a moderate range within a score span of ± 3. This can also be understood from a behavioral finance point of view—highly correlated trading behavior is only expected during times of extreme swing in investors' sentiment.

The right columns in Figure 10 shows the relationships between return and order flow commonality with respect to the attention index. The top figure shows the scatter plot of return commonality *vs* attention, the middle figure shows the scatter plot of order flow commonality *vs* attention, while the lower figure aggregates the attention index into quintiles, and plot them against the mean of the corresponding attention index. Note that there is a linear relationship

between return and order flow *vis-à-vis* attention index in this case, where higher (lower) return and order flow commonality is associated with higher (lower) attention index. This suggests, intuitively that market behaves in a more concentrated manner when attention is higher, but becomes more diversified when attention is lower. This should be intuitive, since trading activities should be the most correlated when there is a higher level of investors' attention— market participants will be scrambling to execute trades during these periods. On the other hand, when attention is low, order flow commonality should be expected to be low, as trading activities will be more diverse due to the lack of market information.

6. Sentiment-Attention based Trading Strategies

The portfolio managers periodically evaluate the performance of their portfolios by tracking individual performance of assets related to the portfolio. Based on the assessment, portfolio managers will engage in rebalancing of their portfolios. In conventional approach, portfolio strategies are generally a function of the distribution of joint returns and their covariances, with constraints on short-selling, turn-over etc. For an extended discussion on the topic of algorithmic trading, see Velu, Hardy, and Nehren (2020). In this approach, the weight attributed to each stock in the portfolio depends on the mean returns and the elements of the covariance matrix of the returns.

In the recent research on portfolio construction, the relationship between returns and the signals such as asset characteristics or exogenous covariates, because of their predictive power, is being considered and is shown to result in better performance. For example, Kelly et al. (2023) and Firoozye, Tan, and Zohren (2022) consider model-based portfolio construction, and these models are closely related to reduced-rank model (3). In Kelly et al. (2023), their criterion is to maximize the covariance between future returns and the signals based on partial least squares—instead of correlation in our reduced-rank regression framework. On the other hand, Firoozye et al. (2022) consider the canonical correlations between return and signals and focus on uncorrelated trading strategies. As we had discussed, both approaches can be unified via the reduced-rank model (3).

In the models we have formulated in this paper, returns and order flows are clearly shown to be associated with investors sentiment and attention. In order to assess the potential benefit to trading strategies given the theoretical advancement made in this paper, we extract a set of portfolio weights[8] resulting from the model (3), and compare their performance to an equal-weighted portfolio. We call our method a "sentiment and attention portfolio strategy". To construct the portfolio weights,, we employ the reduced-rank loadings $\bar{\mathbf{A}} = \mathbf{V}_1' \boldsymbol{\Sigma}_{\epsilon\epsilon}^{-1/2}$ on the 1,213 stocks included in our analysis. We rescale the vector so that $\bar{\mathbf{A}}'\mathbf{1} = 1$. This formulation allows us to investigate the weight on individual stocks to form a market portfolio under the reduced-rank model. On each trading day, we first obtain the attention measure (Tweet volume) of all stocks with a non-zero value on previous day. We trim away the bottom 20% "low attention" stocks as their sentiment measure is less significant. For the remaining stocks, we run the reduced-rank regression model (3) between return and sentiment score based on past 20 days (from t_{-20} to t_{-1}) to generate the $\bar{\mathbf{A}}$ loading matrix. We then normalize $\bar{\mathbf{A}}$ such that $\bar{\mathbf{A}}'\mathbf{1} = 1$, and hold this $\bar{\mathbf{A}}$ weighted portfolio for one day on t_0, which gives us an out sample daily return. The same procedure is then repeated for the following day on a rolling basis.

Figure 11 compares the performance of our sentiment attention portfolio to an equal weighted portfolio. Recall that in the preceding sections, we have demonstrated that the $\bar{\mathbf{A}}$ loading matrix can be used to compute an index to measure market-level return commonality via the operation $\bar{\mathbf{A}}\mathbf{r}_t$. When $\bar{\mathbf{A}}$ is used as portfolio weights, outperformance happens when the $\bar{\mathbf{A}}$ loading vector correctly overweight higher return stocks and underweight lower return stocks. Further research is necessary to fine-tune the construction method for robustness and optimal performance. We should do sparseness on $\bar{\mathbf{A}}$.

7. Conclusions

We have performed a comprehensive longitudinal study of commonality using reduced-rank regression that encompasses other previously used methods such as canonical correlations and principal components. This paper contributes to the existing literature in two main ways. First,

[8]There are now many firms, such as Blackthorne Capital, exploiting the sentiment data.

our analysis is based on a large sample size of 1,213 stocks over an extensive time period of nine calendar years. Our results establish the persistence of commonality. Second, we make use of higher-frequency intraday level sentiment and attention data, another source of commonality, and provide conclusive evidence demonstrating their relationships to returns and order flows.

Our longitudinal study established that commonality in returns and order flows are persistent and remains stable for an extended period of time, but they are merely contemporaneous measures, suggesting that they both driven by exogenous economic or behavioral factors. We model both returns and order flows as endogenous factors, driven by exogenous investors' sentiment and attention. Our reduced-rank regression analyses reveal a nonlinear yet intuitive relationship between commonality in sentiment and attention—when investors' attention on a group of stocks is high, the sentiment they attach to the stocks is either very positive or very negative, since strong sentiment is expected with high attention. When investors' attention on a group of stocks is low, the sentiment they attach to the stocks is in general relatively moderate, this neutral sentiment being the very reason the stocks do not attract their attention, the attention being a scarce resource. This nonlinear relationship also manifests in their impact on returns and order flows.

We finish off our discussion with an exposition on the formulation of a "sentiment and attention portfolio". These analyses not only extend the knowledge of commonality in asset characteristics, but also open up new possibilities to optimize investment or trading strategies. Future work includes introducing sparseness conditions formulated by Witten, Tibshirani, and Hastie (2009), Yuan and Lin (2006), and Zou, Hastie, and Tibshirani (2006) to further trim the $\bar{\mathbf{A}}$ and $\mathbf{B}$ matrices for variable selection. oThis we hope will lead to an elegant interpretation of the reduced-rank regression model.

REFERENCES

Adrian, T., Crump, R. K., Vogt, E., 2019. Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. The Journal of Finance 74, 1931–1973.

Aït-Sahalia, Y., Xiu, D., 2019. Principal component analysis of high-frequency data. Journal of the American Statistical Association 114, 287–303.

Antweiler, W., Frank, M. Z., 2004. Is all that talk just noise? the information content of internet stock message boards. The Journal of Finance 59, 1259–1294.

Baker, M., Wurgler, J., 2006. Investor sentiment and the cross-section of stock returns. The Journal of Finance 61, 1645–1680.

Barberis, N., Shleifer, A., Wurgler, J., 2005. Comovement. Journal of Financial Economics 75, 283–317.

Boehmer, E., Li, D., Saar, G., 2018. The competitive landscape of high-frequency trading firms. The Review of Financial Studies 31, 2227–2276.

Brockman, P., Chung, D. Y., Snow, N. M., 2023. Search-based peer groups and commonality in liquidity. Review of Finance 27, 33–77.

Calomiris, C. W., Mamaysky, H., 2019. How news and its context drive risk and returns around the world. Journal of Financial Economics 133, 299–336.

Chen, J., Tang, G., Yao, J., Zhou, G., 2022. Investor attention and stock returns. Journal of Financial and Quantitative Analysis 57, 455–484.

Chen, X., He, W., Tao, L., Yu, J., 2023. Attention and underreaction-related anomalies. Management Science 69, 636–659.

Cheon, Y.-H., Lee, K.-L., 2018. Maxing out globally: individualism, investor attention, and the cross section of expected stock returns. Management Science 64, 5807–5831.

Chordia, T., Roll, R., Subrahmanyam, A., 2000. Commonality in liquidity. Journal of Financial Economics 56, 3–28.

Chu, L., He, X.-Z., Li, K., Tu, J., 2022. Investor sentiment and paradigm shifts in equity. Management Science 68, 4301–4325.

Cookson, J. A., Engelberg, J. E., Mullins, W., 2023. Echo chambers. The Review of Financial Studies 36, 450–500.

Cookson, J. A., Lu, R., Mullins, W., Niessner, M., 2022. The social signal. Available at SSRN 4241505 .

Corwin, S. A., Lipson, M. L., 2011. Order characteristics and the sources of commonality in prices and liquidity. Journal of Financial Markets 14, 47–81.

Coughenour, J. F., Saad, M. M., 2004. Common market makers and commonality in liquidity. Journal of Financial Economics 73, 37–69.

Cziraki, P., Mondria, J., Wu, T., 2021. Asymmetric attention and stock returns. Management Science 67, 48–71.

Das, S. R., Chen, M. Y., 2007. Yahoo! for amazon: Sentiment extraction from small talk on the web. Management Science 53, 1375–1388.

DeMiguel, V., Nogales, F. J., Uppal, R., 2014. Stock return serial dependence and out-of-sample portfolio performance. The Review of Financial Studies 27, 1031–1073.

Drake, M. S., Jennings, J., Roulstone, D. T., Thornock, J. R., 2017. The comovement of investor attention. Management Science 63, 2847–2867.

Engle, R. F., Kozicki, S., 1993. Testing for common features. Journal of Business & Economic Statistics 11, 369–380.

Fang, L., Peress, J., 2009. Media coverage and the cross-section of stock returns. The Journal of Finance 64, 2023–2052.

Firoozye, N., Tan, V., Zohren, S., 2022. Canonical portfolios: optimal asset and signal combination. Working Paper .

Green, T. C., Hwang, B.-H., 2009. Price-based return comovement. Journal of Financial Economics 93, 37–50.

Han, Y., Huang, D., Huang, D., Zhou, G., 2022. Expected return, volume, and mispricing. Journal of Financial Economics 143, 1295–1315.

Harford, J., Kaul, A., 2005. Correlated order flow: Pervasiveness, sources, and pricing effects. Journal of Financial and Quantitative Analysis 40, 29–55.

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

He, A., Huang, D., Li, J., Zhou, G., 2022. Shrinking factor dimension: A reduced-rank approach. Available at SSRN 3205697 .

Huang, S., Huang, Y., Lin, T.-C., 2019. Attention allocation and return co-movement: Evidence from repeated natural experiments. Journal of Financial Economics 132, 369–383.

Jiang, F., Lee, J., Martin, X., Zhou, G., 2019. Manager sentiment and stock returns. Journal of Financial Economics 132, 126–149.

Jiang, L., Liu, J., Peng, L., Wang, B., 2022. Investor attention and asset pricing anomalies. Review of Finance 26, 563–593.

Kelly, B., Malamud, S., Pedersen, L., 2023. Principal portfolios. Journal of Finance 78, 347–387.

Koch, A., Ruenzi, S., Starks, L., 2016. Commonality in liquidity: a demand-side explanation. The Review of Financial Studies 29, 1943–1974.

Krzanowski, W., 1979. Between-groups comparison of principal components. Journal of the American Statistical Association 74, 703–707.

Kumar, A., Lee, C. M., 2006. Retail investor sentiment and return comovements. The Journal of Finance 61, 2451–2486.

Langlois, H., 2020. Measuring skewness premia. Journal of Financial Economics 135, 399–424.

Lee, C. M., Ready, M. J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46, 733–746.

Longin, F., Solnik, B., 2002. Extreme correlation of international equity markets. Journal of Finance 56, 649–676.

Malceniece, L., Malcenieks, K., Putniņš, T. J., 2019. High frequency trading and comovement in financial markets. Journal of Financial Economics 134, 381–399.

Peng, L., Xiong, W., 2006. Investor attention, overconfidence and category learning. Journal of Financial Economics 80, 563–602.

Reinsel, G. C., Velu, R., Chen, K., 2022. Multivariate reduced-rank regression: theory, methods and applications. Springer New York, NY.

Tauchen, G. E., Pitts, M., 1983. The price variability-volume relationship on speculative markets. Econometrica: Journal of the Econometric Society pp. 485–505.

Vahid, F., Engle, R. F., 1993. Common trends and common cycles. Journal of Applied Econometrics pp. 341–360.

Velu, R., Hardy, M., Nehren, D., 2020. Algorithmic Trading and Quantitative Strategies. CRC Press.

Velu, R., Zhou, G., 1999. Testing multi-beta asset pricing models. Journal of Empirical Finance 6, 219–241.

Wan, R., Li, Y., Lu, W., Song, R., 2023. Mining the factor zoo: estimation of latent factor models with sufficient proxies. Journal of Econometrics Available online: https://doi.org/10.1016/j.jeconom.2022.08.013.

Witten, D. M., Tibshirani, R., Hastie, T., 2009. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics 10, 515–534.

Wood, R. A., McInish, T. H., Ord, J. K., 1985. An investigation of transactions data for nyse stocks. Journal of Finance 40, 723–739.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

Zhou, G., 2018. Measuring investor sentiment. Annual Review of Financial Economics 10, 239–259.

Zou, H., Hastie, T., Tibshirani, R., 2006. Sparse principal component analysis. Journal of computational and graphical statistics 15, 265–286.

Table 1 Descriptive Statistics of Intraday Variables

Year	Sample Stock Return		Order flow		S&P 500		Sentiment	
	Average (bps)	Std Dev *100(%)	Signed Trades	Number of Trades	Average (bps)	Std Dev *100(%)	VIX	Tweet Volume
2012	8.73	2.12	-2.4	1380.1	5.36	0.80	17.80	3977
2013	15.50	1.99	-0.9	1244.2	10.54	0.70	14.23	4953
2014	4.28	1.98	-0.9	1527.6	4.54	0.72	14.18	3689
2015	-0.37	2.17	-7.1	1643.7	0.19	0.98	16.67	4994
2016	10.97	2.36	-1.5	1762.0	3.95	0.82	15.83	7182
2017	7.03	1.99	-2.8	1593.7	7.16	0.42	11.09	7008
2018	-3.48	2.31	-12.9	1883.4	-1.99	1.07	16.64	5862
2019	10.01	2.29	-11.3	1855.8	10.38	0.79	15.39	5569
2020	13.28	4.39	14.5	2766.3	8.32	2.17	29.25	7207

The descriptive statistics of the return, order flow, sentiment, and attention measures of our data sample, summarized over nine calendar years included in our analysis. We use the log midpoint quotes to measure return, and both signed and unsigned number of trades to measure order flow. For comparison, we also include the and the S&P 500 and VIX indexes in the table. Overall, the portfolio of sample stocks exhibit higher average return compared to the S&P 500 index, although its volatility is also higher. The number of trades has increased over time, which is likely due to the increasing prevalence of algorithmic trading. The tweet volume varies over time, but in general exhibits an increasing trend, which may be due to the growth of Twitter over the years. Sample period is from Jan 2012 to Dec 2020.

To add a row for the entire decade at the bottom of the table.

Table 2 Principal Component Analysis for key measures

Measure	First PC	Second PC	Third PC	Cum. Var. Explained
Quote based Return	19.43%	1.91%	1.46%	22.81%
Number of Trades	18.84%	3.08%	1.69%	23.62%
Signed Trades	2.45%	0.69%	0.54%	3.68%
Raw Sentiment Scores	3.09%	1.08%	0.81%	4.97%
Sentiment Z-Scores	2.72%	1.00%	0.72%	4.44%
Tweet Volume	10.11%	2.35%	1.97%	14.42%
Tweet Buzz Measure	4.58%	1.98%	1.45%	8.02%

In general the cumulative variances of return and number of trades are higher than that of sentiment and attention measures. Although the tweet volume and buzz index exhibit higher cumulative variance, we attribute these mainly to the greater use of Twitter over time.

Table 3 anonical Correlation Analysis

Combinations	Contemporary			Lag-1			Lag-12 (Three hours)		
	First	Second	Third	First	Second	Third	First	Second	Third
Ret VS Trade	0.879	0.750	0.738	0.584	0.519	0.480	0.574	0.511	0.433
Return VS Sent	0.849	0.740	0.679	0.848	0.739	0.678	0.847	0.738	0.676
Return VS TVolume	0.878	0.748	0.707	0.878	0.748	0.706	0.877	0.746	0.704
Return VS TBuzz	0.876	0.738	0.687	0.876	0.737	0.687	0.876	0.737	0.685
Trade VS Sent	0.763	0.690	0.670	0.762	0.689	0.669	0.755	0.682	0.655
Trade VS TVolume	0.819	0.757	0.735	0.819	0.756	0.736	0.817	0.754	0.727
Trade VS TBuzz	0.789	0.724	0.700	0.789	0.724	0.701	0.785	0.718	0.692

"Ret" represents the log quote-based midpoint returns, "Trade" is the signed trades measure, "Sent" is tweet sentiment raw score, "TVolume" is the tweet volume, and "TBuzz" is a tweet buzz measure designed by SMA. All of these measures have been standardized to remove the time-of-day effects documented in Wood et al. (1985). Although contemporaneous return and order flow exhibit a high CCA coefficient of 0.879, this relationship falls steeply to 0.584 for the CCA between return and lag-1 order flow. On the other hand, the CCA between return *vs* sentiment and return *vs* attention measures remain high in the range of 0.85, even after introducing multi-period lags. Generally the impact of sentiment scores appears to be higher on return than on trade. Sample period is from Jan 2012 to Dec 2020.

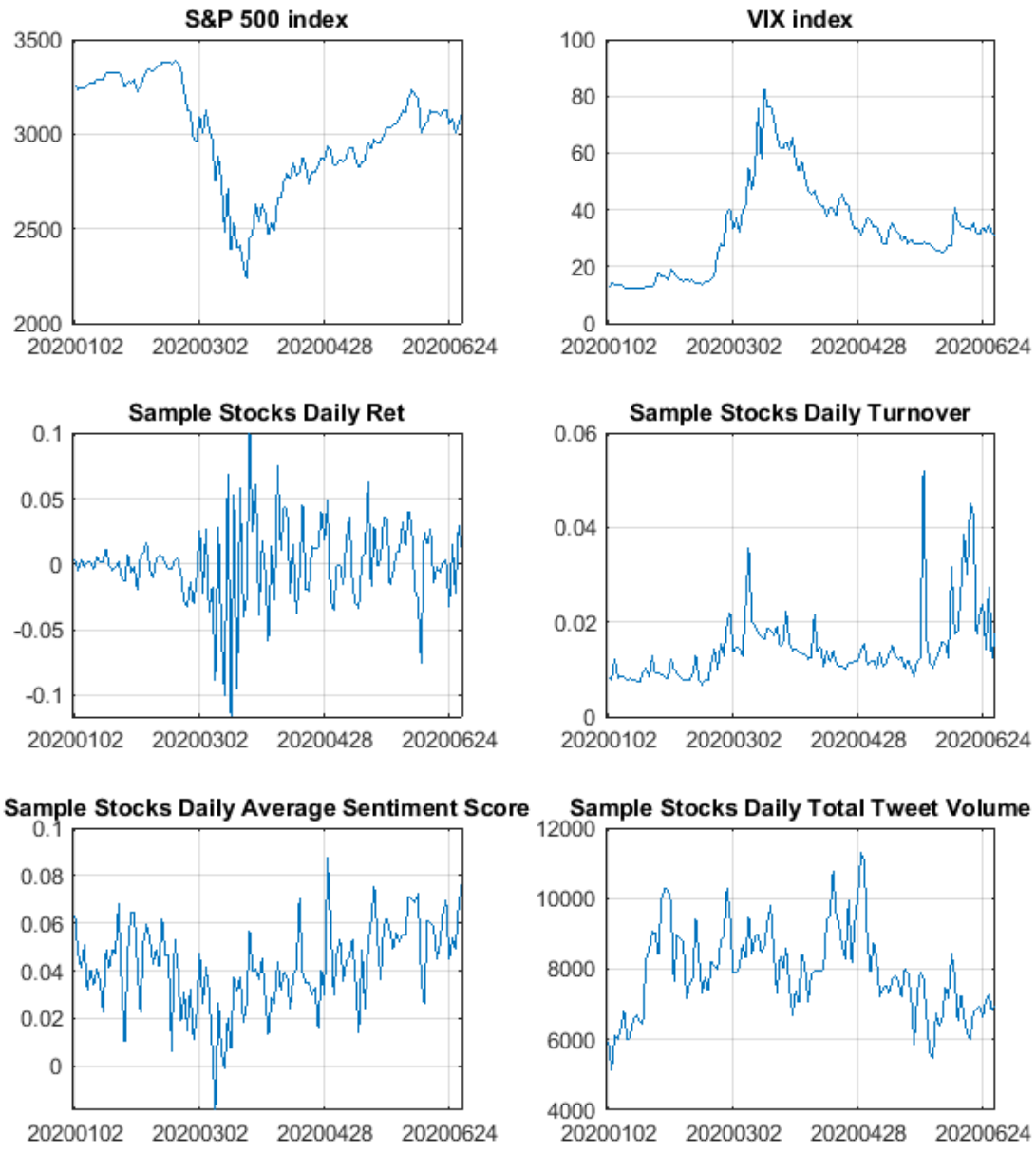

Figure 1. Sample stock measures over the 2020 Covid pandamic period.

The data from Jan-2020 to Jun-2020 at the daily level, a period that included the 2020 stock market crash, when the financial market experienced a spike in volatility, where a substantial crash initially was followed by a quick recovery. During this period, the sample stocks average return is highly correlated with S&P 500 index return (0.93) and with average sentiment score (0.43). The VIX index is negatively related to average sentiment score (-0.38) and positively related to tweet volume (0.24).
Are we covering both the crash and the pandemic here?

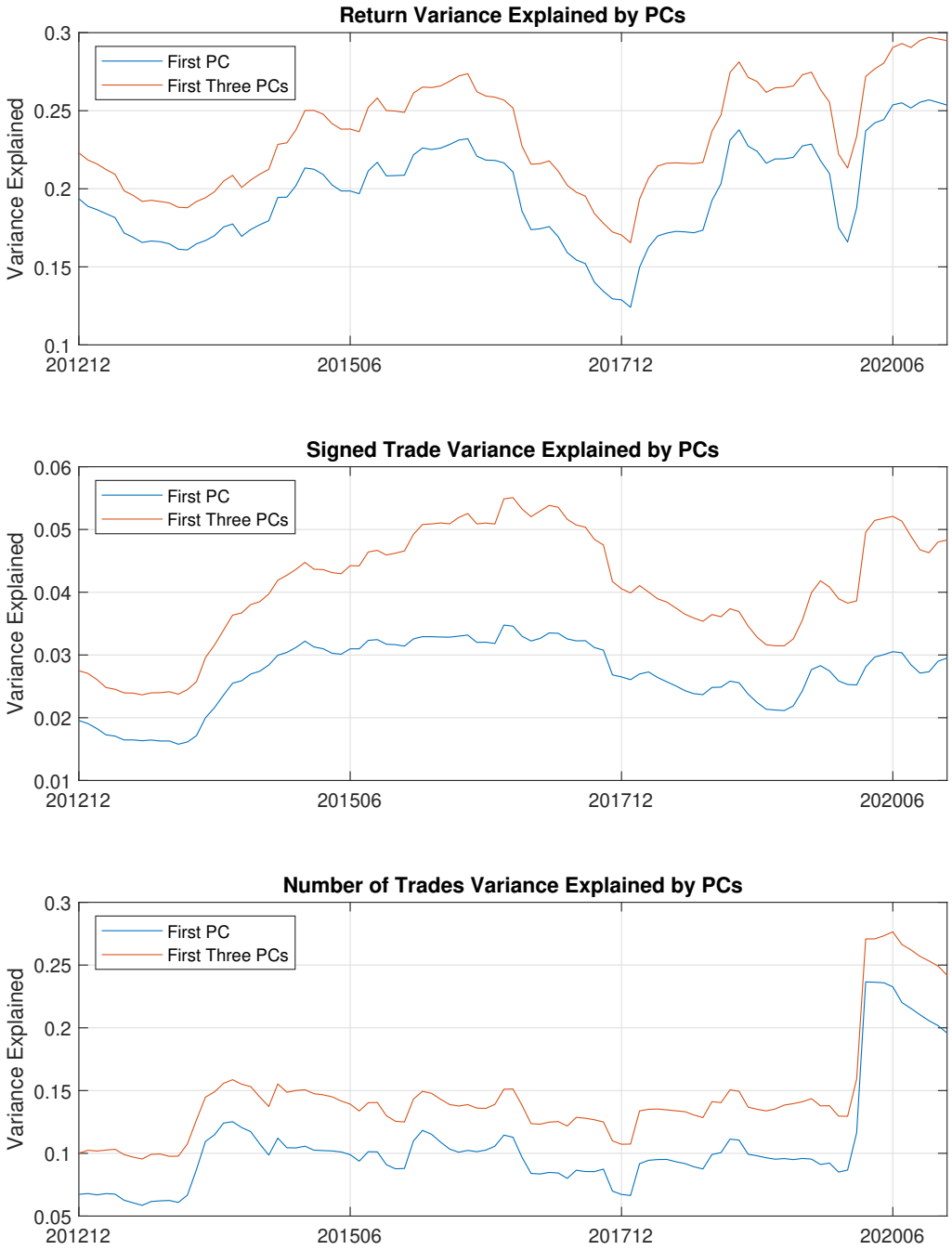

Figure 2. The principal component analysis on a 12-month moving window for return (top) and both signed (middle) and unsigned trade (bottom) measures, showing the longitudinal variation in returns and order flows commonality over the period Dec 2012 through to Dec 2020. The variance explained by the dominant PCs is often used as a commonality proxy. This figure suggests that longitudinally commonality has been trending upward, with a spike experienced during the March 2020 Covid crash. Comment on the dip in 2017. Shorted figure title: (1) Returns, (2) Signed Trades, (3) Number of Trades.

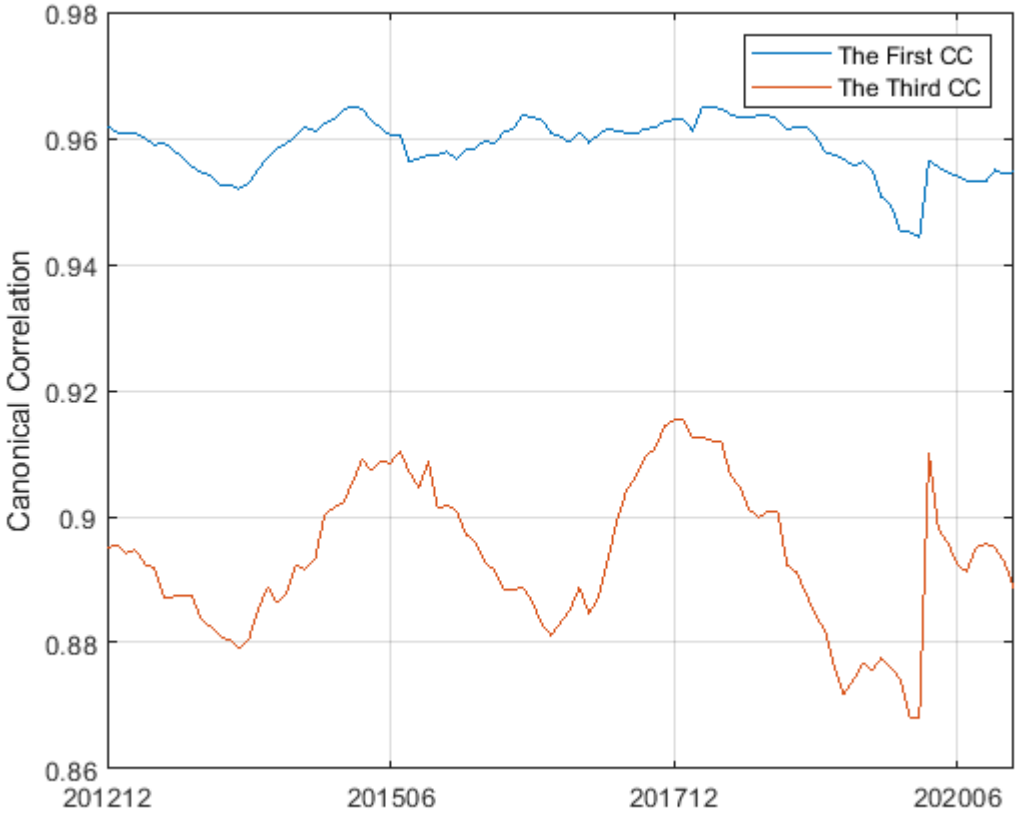

Figure 3. 12-month rolling Canonical Correlation between returns and order flows. The variations are also mild, suggesting a stable relationship between returns and order flows over the decade. The high canonical correlations indicate that returns and order flows have strong co-movement and are likely to be driven by some common factors. Add Title: Returns and Orderflows

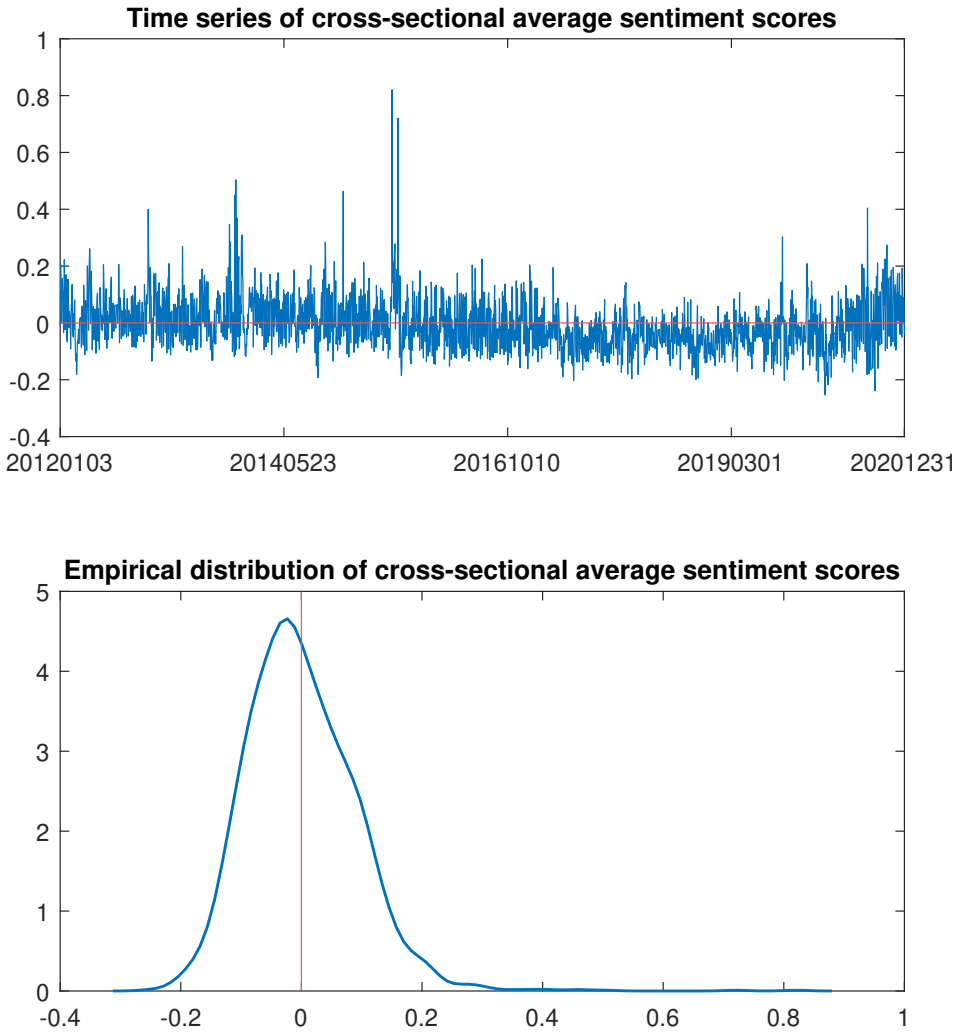

Figure 4. Cross-sectional mean of sentiment scores after removing time-of-day effect (2012.1-2020.12). The time series plot is provided in the upper figure, while the bottom figure shows the empirical distribution. Both plots show an evident right-skew in the sentiment scores' distribution, suggesting that positive sentiments has a larger variation, ranging from mildly positive to exuberant, while the variation in negative sentiments is more muted. Title (l) replace "of" with "plot of". Add measures of autocorrelation on the scores and on squared scores.

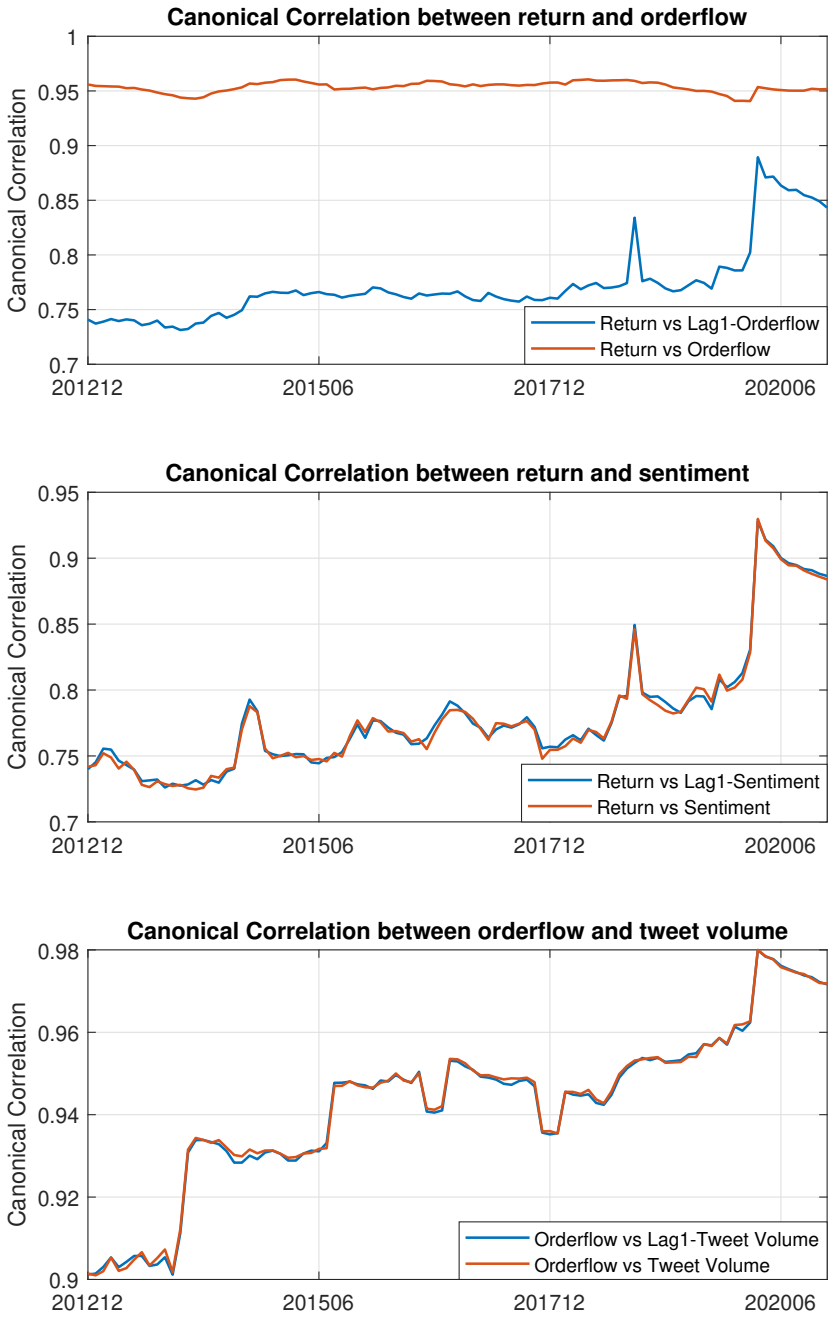

Figure 5. 12-month rolling canonical correlations (CC) among return, order flow, sentiment and attention. The CC between return and order flow is on average 0.95, but between return and order flow with lag-1 dropped to approximately 0.77. On the other hand, the CC between return and sentiment (as well as between order flow and attention) almost does not vary when introducing a lag. Title and axis label update.

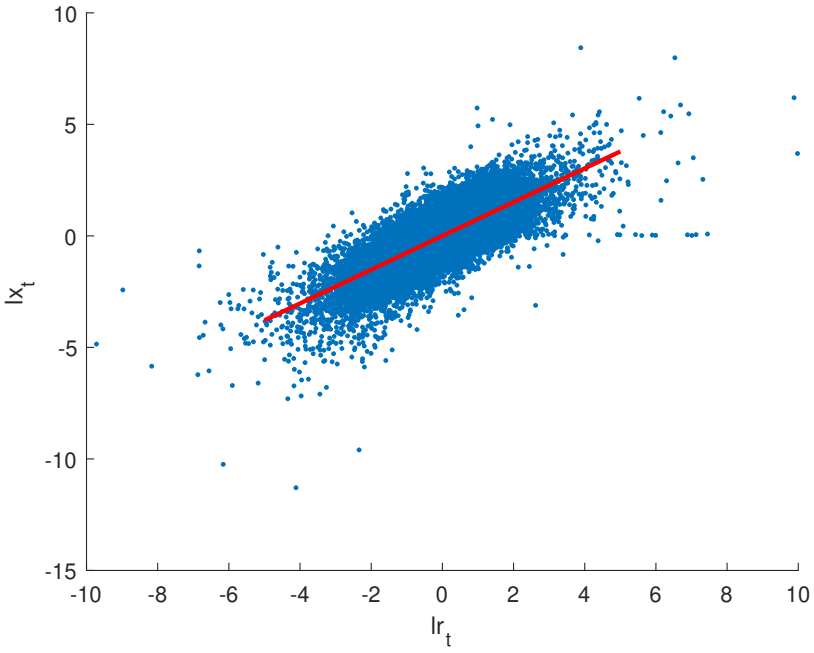

Figure 6. Canonical scores based on the smallest canonical correlation in model (5). We use the smallest canonical correlation between $\mathbf{r_t} - \mathbf{x_t}$ and $\mathbf{x_t}$ to construct linear combination of $\mathbf{r_t} - \mathbf{x_t}$ following the relationship $l_i'\mathbf{C} = 0$ that is independent of $\mathbf{x_t}$, so that $l'\mathbf{r_t} \sim l'\mathbf{x_t}$. $\mathbf{x_t}$ has been rescaled so that $\mathbf{x_t}$ and $\mathbf{r_t}$ share the same covariance matrix. This figure suggests it is possible that both returns and order flows have common sources of influence. The R^2 of the univariate regression (red line) is 0.5732. Sample period covers from Jan 2012 through Dec 2020. Title, x-label and y-label update.

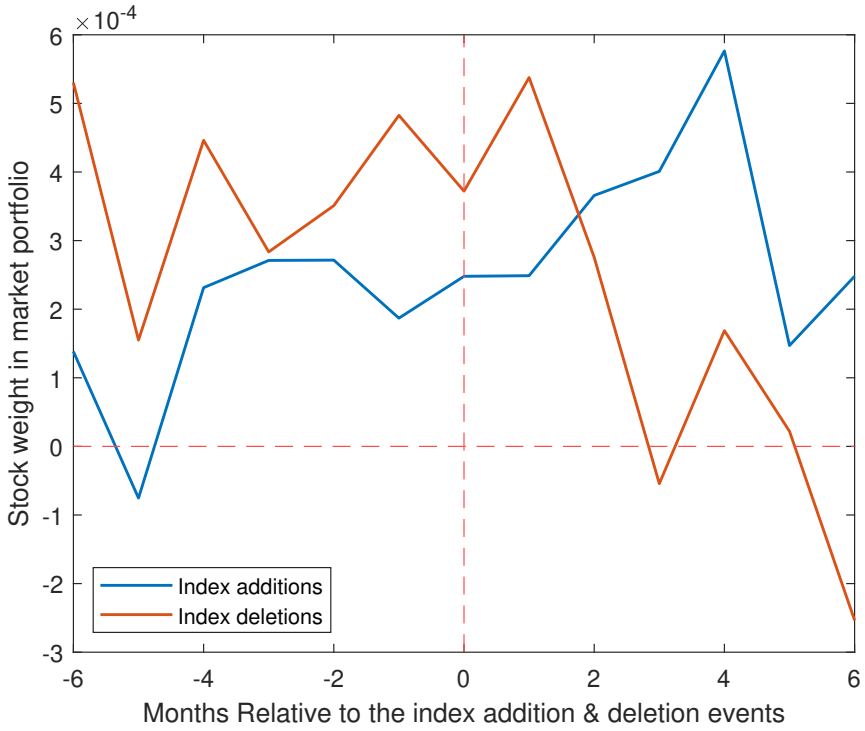

Figure 7. Average reduced-rank stock weight in the market portfolio around the event period of addition to or removal from the S&P 500 index constituent. The stronger liquidity demand for stocks in the S&P constituent leads to a higher return and order flow commonality, manifesting in a higher (lower) weight for constituent (non-constituent) stocks. The result of the event study here clearly demonstrates a reduced-rank weight increase (decrease) for stocks inclusion (exclusion) to the S&P 500 index. To include a graph for the rest of the stocks here.

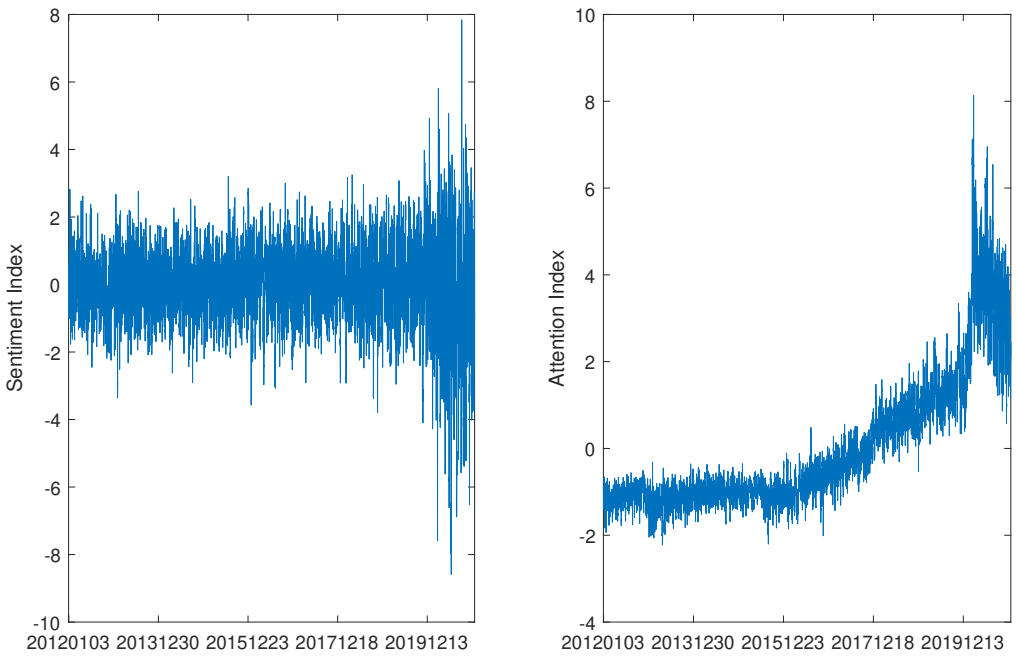

Figure 8. Time series plots of sentiment and attention index. Investors' attention reaches the peak during pandemic period. The correlation between our sentiment and attention indices is -0.256. We fit a GARCH(1,1) model for the sentiment index and an ARCH(1) model for the attention index—the models confirm that both are non-stationary. Add plots of returns/order flows as well to this figure.

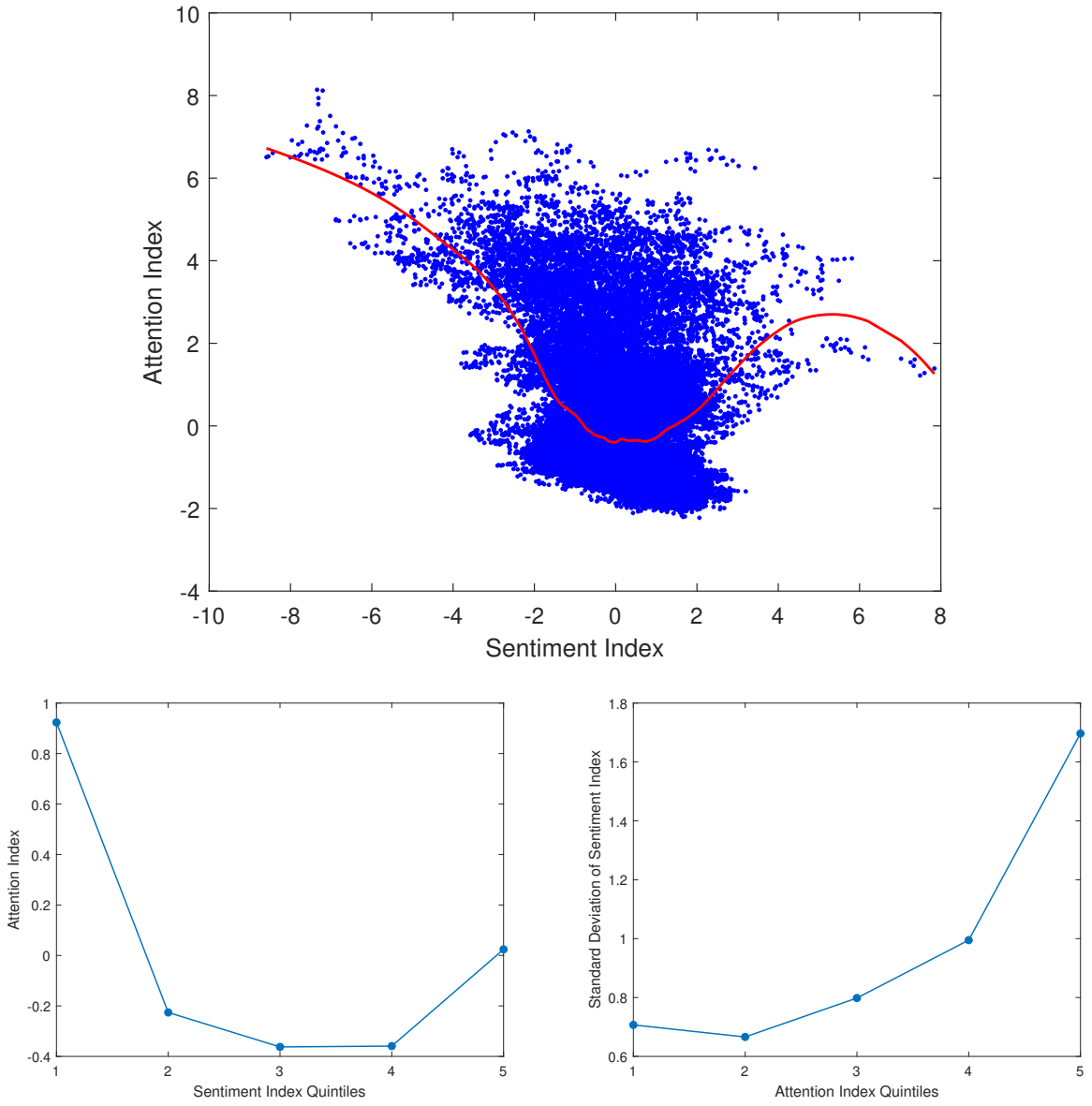

Figure 9. Scatter (top) & quintile (bottom) plots of attention index *vs* sentiment index. The scatter plot clearly indicates that high attention is associated with strong sentiments, i.e. very positive or very negative sentiment. On the other hand, when the sentiment index is closer to 0, attention can vary from low to high. This suggests a nonlinear relationship between attention and sentiment, which is graphically illustrated in the quintile plot of the average indices in the bottom left figure. On the other hand, when attention is low, the dispersion of sentiment index, measured by its standard deviation is low, while high attention is associated with large dispersion in the sentiment index, as shown in the bottom right figure.

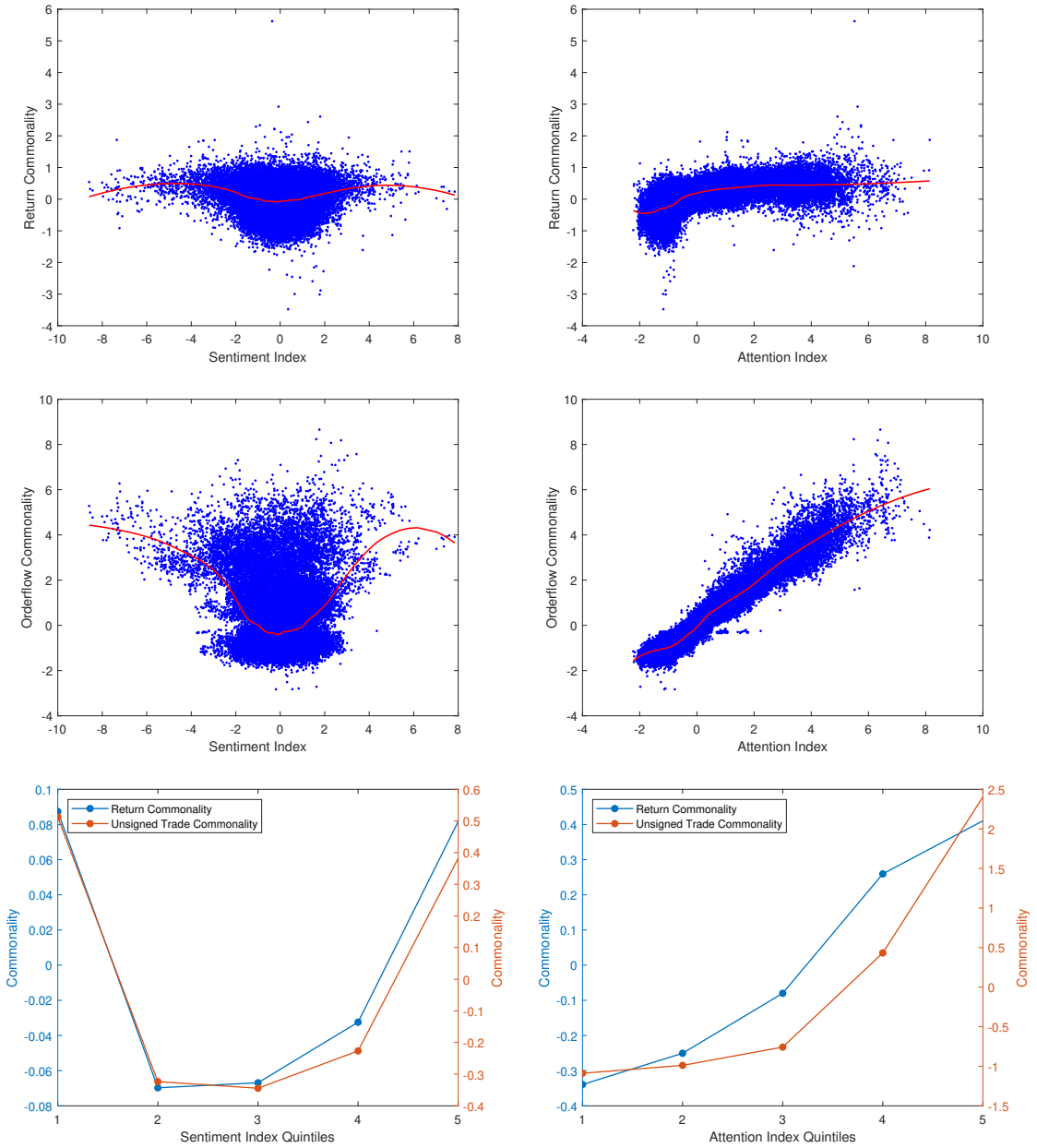

Figure 10. Relationships between return and order flow commonality with respect to the exogenous sentiment (left) and attention (right) indices. Note the distinct nonlinear relationship on the left figures—return and order flow commonality are high when the sentiment is strong (very negative or positive), but low when the sentiment is muted. On the other hand, the right figures show that return and order flow commonality are linearly associated with attention index—high attention leads to strong commonality, and vice versa.

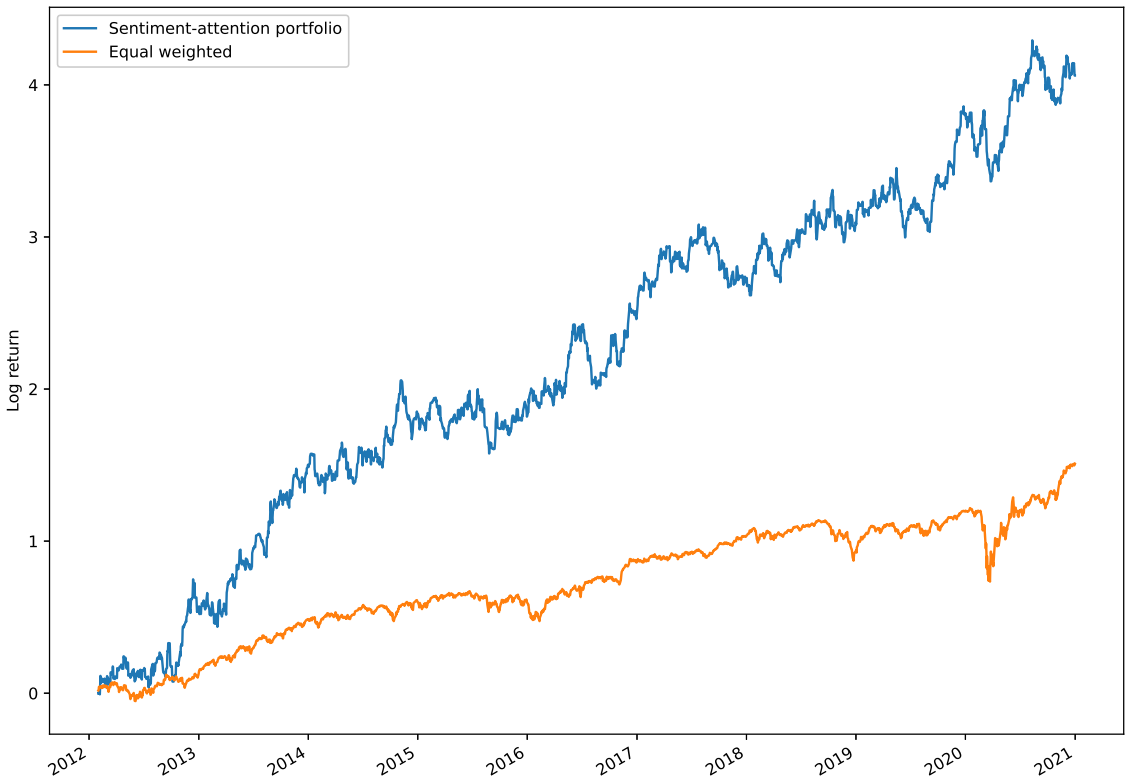

Figure 11. The sentiment and attention portfolio constructed using our reduced rank regression model (3), compared to the performance of an equal weighted portfolio. Significant outperformance is obtained by filtering for high attention stocks, and using investor sentiment as exogenous factor to construct the $\bar{\mathbf{A}}$ loading vector to overweight higher return stocks and underweight lower return stocks.