3-2020

Why commonality persists?

Raja VELU

Zhaoque ZHOU

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

Citation

VELU, Raja; ZHOU, Zhaoque; and TEE, Chyng Wen. Why commonality persists?. (2020). 1-46.
Available at: https://ink.library.smu.edu.sg/lkcsb_research/6569

Why Commonality Persists?

Raja Velu[*] Zhaoque Zhou[*] Chyng Wen Tee[†]

Current Version: February 2020

ABSTRACT

Studies on commonality in returns, order flows and liquidity find that the first principal component is closely aligned with the market factor. With the increasing presence of high-frequency trading, commonality in returns, order flows, and liquidity can potentially arise from the commonality in the interpretation of real-time signals. In this paper, we go beyond the first factor and show that the other dominant principal components consistently reflects investors' herding behavior, demonstrating the multi-dimensional aspect of commonality. Instead of relating the asset returns to order flows, we take both as endogenous, and provide empirical evidence showing that returns commonality is driven by investors' attention, while order flows commonality is driven by investors' sentiment. We also present a comprehensive longitudinal study of commonality and co-movement over a period in excess of two decades under a unifying market microstructure framework to demonstrate the persistence of commonality over time. Our results not only extend the knowledge about cross-sectional asset behaviors, but can also be used to develop systematic trading strategies.

Keywords: *commonality; order flow and liquidity measures; co-movement; market sentiments; investor attention; principal component analysis; canonical correlation analysis; reduced rank regression.*

[*] Department of Finance, Whitman School of Management, Syracuse University, Syracuse, New York 13244.

[†] Lee Kong Chian School of Business, Singapore Management University, 50 Stamford Road #05-01, Singapore 178899

1. Introduction

A central thesis of Markowitz's portfolio theory is that investors aim to maximize return for a specified risk level, and that risk can be reduced by holding a diversified portfolio of assets. However, many studies have since observed that there exists a strong commonality in stock returns, order flows and liquidity. This is antithetical to what is advocated by the portfolio theory. Traditional economic theory also states that the price of a stock is only expected to vary if there is a change in its fundamental value. If this reasoning is correct, then the commonality in returns can only arise when there is a commonality in fundamental values. Nevertheless, this assumes that there are no frictions and market consists of rational investors. Given that these assumptions rarely hold in practice, commonality can arise due to factors beyond fundamental values.

Earlier pioneering work on the study of assets' cross-sectional behaviors by Chordia, Roll, and Subrahmanyam (2000) and Hasbrouck and Seppi (2001) has demonstrated the existence of commonality in returns, order flows, and liquidity measures. In addition, commonality in order flows has been postulated as the source of commonality in returns. Subsequent studies have focused on identifying sources of commonality in order flows—these research can be broadly classified as providing "supply-side" or "demand-side" explanations, with specialists and market makers representing the supply-side, while institutional traders and fund managers alike representing the demand-side. Coughenour and Saad (2004) present a supply-side explanation whereby commonality arises through the trading patterns of specialists who provide liquidity for multiple stocks. On the other hand, Koch, Ruenzi, and Starks (2016) give a demand-side explanation by focusing on mutual funds. They argue that mutual funds generally hold large portfolios, and the net flows due to liquidity shocks are likely to be correlated across funds.

In our view, it is unlikely that supply or demand side can operate in isolation. Instead, we argue that both returns and order flows should be taken as endogenous, and are driven by exogenous, non-fundamental, and behavioral factors. Barberis, Shleifer, and Wurgler (2005) have pointed out that behavioral and non-fundamental factors can come from various sources, most of which is related to the way investors process asset-related information. Among them is the habitat view, which postulates that most investors consider only a small set of assets for investment. This consideration set is formed based on the investors' past experience that may be related to available information or “sentiments” attached to certain assets. Another view is driven by the argument that investors’ “attention” is a scarce cognitive resource, and is based on how they process information at the macro level—most investors will group assets into broad categories and allocate their investments over these categories, without paying too much attention to the individual asset. The overlap among investors’ sentiment and attention will result in correlated behavior, thus manifesting as commonality in returns, order flows, and liquidity.

Classical finance theory leaves no role for investor sentiment or attention, and assumes that market participants are rational investors. Recent research on behavioral finance has nevertheless refuted this view. For instance, Baker and Wurgler (2006) uses a collection of publicly observable empirical measures to proxy investors’ sentiment, and demonstrate that it can be related to future stock returns. A related stream of research focuses on investor attention (see Peng and Xiong (2006)), which suggests that limited attention leads investors to focus on market- and sector-wide information more than on firm-specific information, implying a link between investor attention and market returns. Chen, Tang, Yao, and Zhou (2019) show that individual investor attention proxies proposed in the literature collectively have a common component that has significant power in predicting stock market excess returns.

With the increasing speed of information diffusion and the presence of algorithmic traders, investors’ sentiment and attention can be taken as the exogenous factors driving the commonality in returns and order flows. This information diffusion view has been observed during the time of earnings announcements in particular. Malceniece, Malcenieks, and Putniņš (2019) perform a rigorous investigation into the aspects of co-movement due to friction and high frequency trading.

In this paper, we provide empirical evidence to show that commonality in returns are linked to investors’ attention, while commonality in order flows are linked to investors’ sentiment, thus demonstrating the multi-dimensional aspect of commonality. We also formulate a unifying microstructure model, and perform the first longitudinal study of commonality over an extended period of 25 calendar years to investigate its persistence over time. Our methodologies include a combination of reduced-rank regression, principal component analysis, and canonical correlation analysis. Our analysis clearly indicate that the first principal component can be interpreted as the well-known market factor, while the subsequent dominant principal components can be used to measure investors’ herding behavior, and is closely associated to investors sentiment and attention at both the low and high frequency levels.

This paper is organized as follows: we begin with an exposition of our modeling framework in Section 2, followed by a brief description of the data set used in our study in Section 3, including the high-frequency Twitter feed used in our sentiment & attention analysis. The longitudinal study is presented in Section 4, while the analysis on the sources of commonality is covered in Section 5. Finally, conclusions are drawn in Section 6.

2. Modeling Framework

There are two common established methodologies used for studying commonality in the literature, namely regression analysis and Principal Component Analysis (PCA). In the regression analysis approach, the individual asset's characteristic (returns, order flows, or liquidity measures) is regressed on a market level composite metric, and the significance of the slope coefficients is used to infer commonality (see Chordia et al. (2000)). In the PCA approach, after standardization of the variables over individual assets (and thus forming a correlation matrix), the principal components are extracted. The amount 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. It is also hypothesized that because order flows contain informed components, they are likely to explain the commonality in returns. This analysis is done via canonical correlation analysis.

Identifying the common features among financial time series has been the focus of numerous studies, including Engle and Kozicki (1993) as well as Vahid and Engle (1993) among others. The time series are studied for their co-movement if they are stationary, and for their co-integration if they are non-stationary. In this paper, we focus on stock returns, order flows and liquidity. We postulate that a set of common and idiosyncratic variables that represent economic fundamentals or behavioral patterns (as measured by investors' attention and sentiment) can result in commonality. Similar forecasts of macroeconomic variables and trading strategies such as momentum can also lead to correlated trading that may result in commonality, along with their association with investors' sentiment and attention.

Economic time series in general share some common characteristics such as trends, seasonality and serial correlations. The existence of common elements is identified through the indicators of co-movement. 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 $m \times 1$ return vector, m is the number of assets, $\mathbf{f_t}$ is a lower dimensional common feature vector. If $\mathbf{f_t}$ is a $r \times 1$ vector, where r is the number of features, then the matrix $\mathbf{A}$ is of dimension $m \times r$. The $\mathbf{f_t}$ vector can either be constructed from the past values of $\mathbf{r_t}$, or it can also be constructed from exogenous variables, say $\mathbf{x_t}$ (the order flow), which is the approach taken by many studies on commonality in the literature. These two can be characterized by either a vector auto-regressive (VAR) model or a multi-variate regression model as follows:

$$\begin{aligned} \text{VAR(1)}: \qquad \mathbf{r_t} &= \mathbf{C}\mathbf{r_{t-1}} + \mathbf{a_t}, && (2)\\ \text{Multivariate Regression}: \qquad \mathbf{r_t} &= \mathbf{C}\mathbf{x_t} + \mathbf{a_t}. && (3) \end{aligned}$$

Note that the coefficient matrix $\mathbf{C}$ is of lower rank in both cases, and thus can be decomposed as,

$$\mathbf{C} = \mathbf{A} \times \mathbf{B}, \tag{4}$$

where $\mathbf{A}$ is a $m \times r$ matrix, $\mathbf{B}$ is a $r \times m$ matrix, and $\mathbf{f_t} = \mathbf{B}\mathbf{r_{t-1}}$ (or $\mathbf{B}\mathbf{x_t}$) are the common factors. For the model in Equation (2), we choose only a VAR(1) model for returns, since returns in general do not carry much memory beyond lag 1. Note that in the model specified by Equation (3), the coefficient matrix contains elements of the market impact of trades on prices. The existence of commonality will be reflected in the cross-impact coefficients, i.e. the off-diagonal elements of the matrix $\mathbf{C}$.

The multivariate regression model in Equation (3) in its general form can also be used to include other variables that might potentially affect commonality. Note that a key theme in the commonality literature is that a relatively small number of factors will suffice in driving the commonality of the larger set. A generalized multivariate regression model that we will use in this paper takes the form:

$$\mathbf{r_t} = \mathbf{C}\mathbf{x_t} + \mathbf{D}\mathbf{z_t} + \mathbf{a_t}. \tag{5}$$

As an example, $\mathbf{z_t}$ could denote investors' sentiment or attention scores, which we will explore in later sections as the potential source of commonality.

The main statistical tools we employ for the model identification are principal component analysis (PCA), canonical correlation analysis (CCA), and reduced rank regression (RRR). In the conventional approach, the liquidity measures at the individual stock level are related to market level measures either via simple regressions or PCA, and commonality is inferred through the strength of the regression coefficients or through the variance explained by the first principal component. To the best of our knowledge, no explicit modeling attempt has been made so far to check if there is any similarity among coefficients over different stocks. This is simply equivalent to building a multivariate regression model of the type (5), which is essentially a set of multiple regressions, with each regression representing a single stock. What binds these equations are constraints over coefficients across stocks.

One of the modeling contribution of our paper is to handle this analysis via Reduced-Rank Regression which accommodates these constraints which are not explicitly known, but implied by the data structure. For example if the factors $\mathbf{f_t}$ are determined by $\mathbf{x_t}$, via

$$\mathbf{f_t} = \mathbf{B}\mathbf{x_t} + \mathbf{u_t}, \tag{6}$$

then

$$\mathbf{r_t} = \mathbf{C}\mathbf{x_t} + \mathbf{a_t} = \mathbf{A}\mathbf{B}\mathbf{x_t} + \mathbf{a_t}, \tag{7}$$

where the matrix $\mathbf{C}$ is of lower rank. The direct relationship between returns and order flows in Equation (7) is more efficient to model than relating the factors from two sets of series as in Hasbrouck and Seppi (2001). The rank of the matrix $\mathbf{C}$ can be taken to indicate the effective number of factors driving the returns and order flow relationships. Full details on the estimation and inference procedure of the model in Equations (7) and (5) under the reduced-rank regression approach are given in Reinsel and Velu (1998). These will be described in more detail when results are presented in subsequent sections.

One might notice some resemblance between the principal components of the returns and the matrix $\mathbf{A}$ in Equation (7). Note if the rank of $\mathbf{C}$ is 1, then from Equation (7), we can write

$$\mathbf{\Sigma_{rr}} = \mathbf{AB\Sigma_{xx}B'A'} + \mathbf{\Sigma_{aa}}, \tag{8}$$

where $\mathbf{A}$ is simply a $m \times 1$ column vector α, and $\mathbf{B}$ is an $1 \times m$ row vector β'. Writing $\mathbf{\Sigma_{rr}}$ in terms of its principal components, $\mathbf{\Sigma_{rr}} = \sum_{i=1}^{m} \lambda_i \mathbf{p_i p_i}'$, one obtains

$$\lambda_1 \mathbf{p_1 p_1}' + \sum_{i=2}^{m} \lambda_i \mathbf{p_i p_i}' = (\beta' \mathbf{\Sigma_{xx}} \beta) \alpha \alpha' + \mathbf{\Sigma_{aa}} \tag{9}$$

By equating the terms, it can be seen that $\mathbf{p_i}$ is proportional to α, highlighting the relationship between the principal component of $\mathbf{r_t}$ and the matrix $\mathbf{A}$.

Finally, we define a direct yet intuitive measure of commonality within the regression framework of Equation (7). Note that the elements of the matrix $\mathbf{C}$ measure both own- and cross-price impacts: the diagonal elements of that matrix measure own-impacts, while the off-diagonals represent the cross-impacts. If there is no commonality, then the matrix $\mathbf{C}$ should be diagonal. An index λ^2 that measures commonality via cross-impacts can therefore be defined as:

$$\lambda^2 = 1 - \left(\frac{\sum_i C_{ii}^2}{\sum_i C_{ii}^2 + \sum_{i \neq j} C_{ij}^2} \right). \tag{10}$$

As the distribution of the estimated matrix $\mathbf{C}$ is known, it is possible to derive the distribution of this index under the null hypothesis that cross-impacts are zeros, or there is no commonality. Our commonality measure complements the existing methodology used widely in the literature, which simply measures the variance contribution of the dominating principal components.

3. Data

3.1. Returns, Order Flows, and Liquidity Measures

The returns and order flows data are obtained from the NYSE's Daily Trade and Quote (TAQ) database. Following Hasbrouck and Seppi (2001), we focus our analysis on thirty large market capitalization stocks at any given time. However, we extend the length of study to cover a significantly longer period (from January 1st, 1994 through to December 31st, 2018) of 25 calendar years, as one of our primary goal is to investigate the longitudinal behavior of commonality over an extended period, during which time the market has gone through numerous turbulent changes of boom and bust. The thirty sample stocks are chosen each year as the constituent stocks of the Dow Jones Industrial Average Index. The Dow sample consists of generally large cap stocks that are liquidly traded, and are largely transacted by institutional traders, and therefore should exhibit more commonality.

We aggregate the intraday tick-level data from TAQ into 15-minute intervals, resulting in 26 observations per trading day. The chosen time resolution of 15-minute will smooth out market frictions and yield better estimates of returns and order flows. Returns are calculated as the difference in the log midpoint quotes over the two endpoints of each interval. For order flow, we use both signed and unsigned measures. We drop the firm index 'i' in the interest of brevity. For the j^{th} trade (where $j = 1, \ 2, \ \cdots$), 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 "buy" trade and $\text{sign}(v_j) = -1$ for "sell" trade, we also construct signed order flow measures. We use the matching algorithm by Lee and Ready (1991) to classify the direction of each trade.

In addition to returns and order flows, many studies have also considered various liquidity measures. This is based on the observation that periods of significant market decline or heightened volatility typically increase the demand for liquidity, as agents liquidate their positions across many assets and reduce the supply of liquidity as capital constraints are hit. The full list of order flows and liquidity variables used in our analysis, along with their definitions, is presented in Table 1. Here, P_j, v_j and $\text{sign}(v_j)$ are the price per share, share volume, and the direction of the j^{th} trade, respectively, all of which are aggregated into 15-min intervals.

3.2. Investor Sentiment and Attention

Financial news can be classified into two main groups: structured and unstructured. Certain structured financial news related to a particular stock are released on a regular basis (such as earnings statements), and models to analyze such events are well studied. On top of that, there are other unstructured news related to the product or personnel of the company that may come unscheduled, and may also affect trading decisions. These are generally termed as news-based event strategies, and have also been fairly well-studied in the finance literature.

Unstructured news streams do not come at regular intervals. They are qualitative, and are usually presented in text format. In order to relate this information to the stock performance, they need to be appropriately quantified. Because this information originates from multiple sources, including social media, it also needs to be properly aggregated, in order to extract the signal from the noise.

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 techniques. For instance, by relating the sentiment to the performance of stocks in the Morgan-Stanley High-Tech Index, they observe that although there is no strong relationship to individual stock prices, there is a statistical relation to the aggregate index performance. Furthermore, a strong relationship between message volume and volatility is also established, consistent with earlier research by Antweiler and Frank (2004). These examples clearly illustrate that investors sentiments expressed via media carry meaningful impact on stock behavior.

Several technology companies (iSentium, SMA, MarketPsych, RavenPack, etc.) have sprung up in the last decade offering the service of aggregating the web information or sentiment related to stocks. Table 2 provides some brief details on a few providers of web analytics. We use SMA for our analysis in this work, the data of which is made available to us for nine calendar years, from 2010 to 2018. The data comprises of sentiment scores for a time stamp adjusted for the historical (20-day) average as well as its volatility. In addition, volume of tweets, along with a measure of unusual buzz in the activity, and a dispersion measure for tweet source diversity contributing to the sentiment score are also provided. The sector and industry classification are tagged.

4. Commonality in returns, order flows, and liquidities

4.1. Descriptives Statistics

Table 3 presents the descriptive statistics of order flow and liquidity measures of our sample stocks across the time period studied. Since the constituent stocks of the Dow index can vary from year to year, the total number of stocks that appear in our 25-year study period is greater than thirty. Statistics for the year 2008 (when the market collapsed) along with the year 2009 (when it recovered) stands out in particular compared to the other years. As expected, the turnover and volatility in these two years experienced a dramatic spike, clearly indicating that these two years exhibited unusual market behavior. We will examine in details how these translate to the commonality measures. Figure 1 presents the results obtained by running the analysis on the data for 2008 and 2009 at the monthly level. It is obvious that much of the variation occurs in the months surrounding October 2008, when the market lost confidence, which is also reflected in the investor sentiment scores based on Baker and Wurgler (2006).

We consider the eight order flow measures listed in Table 1. These measures reflect the total and the imbalance in trading activities between buy and sell orders. By construction, these variables are of different magnitude. In order to formulate the reduced-rank model in Equation (7), we will need to select one variable among these eight that can capture the correlation among the stock universe included in our analysis. Nevertheless, it will be interesting to first study the commonality among these order flow measures. We start by standardizing our variables to have unit variance before performing principal component analysis (PCA) on these measures, the results of which are presented in Figure 2. Pooling the results, it is clear from Figure 2 that there are between two to three dominant components.

Table 4 presents the principal component coefficients for all eight order flow measures between 1994-2006 and 2007-2018. Interestingly the structure of the components also seem to change over time. Before 2006, the first principal component has approximately equal weights to all eight measures, while the second principal component of signed and total measures are of opposite sign. But starting from 2007 onwards, there is a clear separation between the first and second components. The first component is all total (unsigned) measures and the second component is all signed measures. The third component (not presented here), which is a contrast between the sum of signed trade, signed square root of dollar volume and other two signed volume measures, has remained roughly the same.

Our PCA analysis reveals a high degree of correlations among the order flow measures. Given this, 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 (7). However, 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. Using an index measure (such as a principal component) might make it difficult to interpret our results, and hence it is more intuitive to use the single most representative order flow measure that exhibit the highest correlation with the return. Here we choose signed trades because it is generally the most highly correlated with returns at the individual stock level[1].

Table 5 presents the principal component coefficients for six liquidity measures defined earlier. Here "Dollar Depth" is the simple average of the best bid quote and best ask quote in dollars, while "Shares Depth" is the similar definition in shares. The first component can be taken as an average of all liquidity variables except for depth variables. The second component reflects the depth, where both depth variables have equal weight. The third component reflects the contrast between quoted spread measures and effective spread measures.

4.2. Longitudinal Study

In order to investigate the persistence and dimensionality of commonality over time, we extend the results of Hasbrouck and Seppi (2001) to cover a 25-year period from January 1st, 1994 through to December 31st, 2018. We follow the same standardization procedures for all the variables, and the eigenvalues should add up to the number of stocks considered (m). If there is a high correlation among the variables, the proportion of the variance explained by the first principal component should dominate. On the other hand, if the variables are uncorrelated, then we can expect each component to have equal contribution. This rule of thumb provides a way to interpret the magnitude of the eigenvalues that result from the PCA analysis. More formally, under the null hypothesis that there is no commonality, one could use the confidence interval based on the distribution of $\sqrt{T}(\lambda - \Psi) \sim N(0, 2\Psi^2)$ to check for the presence of commonality. This results in a confidence interval for Ψ of $\left(\frac{\lambda}{\sqrt{1+2\tau}}, \frac{\lambda}{\sqrt{1-2\tau}}\right)$, where $\tau^2 = 2/T$. With large T, this interval generally tends to be narrow. In our choice of the number of principal components, we have selected components with their eigenvalue ('λ') greater than 0.7 as recommended in the literature (see Jolliffe (2002), p115), though we report only the first three components to be consistent with the literature on commonality.

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

The longitudinal principal components of returns, order flows, and liquidity measures are presented in Table 7. The first three components generally account for close to 50% of the variation in recent years, with the first component being the dominant contributor. Our results in the early years between 1994 and 1998 are comparable to the results reported in Hasbrouck and Seppi (2001) (who studied one full calendar year of 1994) and in Corwin and Lipson (2011) (who studied four calendar months from November 1997 to February 1998). As one would expect, the commonality index peaked in 2008, when there was a significant market crash. This is also reflected in the signed trade and Amihud's liquidity measures as well, all of which are plotted in Figure 3. The commonality in return is higher than the commonality in other measures, though they all exhibit comparable trend and variation over the period studied.

We also investigate the diurnal aspect of the commonality behavior. The patterns of commonality as observed in Harford and Kaul (2005) over certain durations of a trading day seem to be different, especially during the last trading hour of the day. They point out that if correlated trading effects are driven by institutional transaction, and institutions prefer to trade near the end of the day, these effects should be most apparent at the end of each day. We analyze our dataset and obtain consistent results. The diurnal patterns of commonality in returns, order flows, and liquidity are clearly visible in Table 6, which presents the largest eigenvalue from the PCA for all three variables, contrasting the last trading hour with the rest of the trading hours. We also observe that while the effect is true for all the years in the study (1994-2018), it is dramatically more pronounced in 2008, as one would expect when the market is under stress.

4.3. Canonical Correlations and Multivariate Regression

The commonality in returns and in order flows are clearly established. If they are statistically correlated, a direct way to study this correlation is through canonical correlation analysis, as expounded in Hasbrouck and Seppi (2001). These correlations capture the best linear relationship among the two sets of variables. Recall that the price impact model in Equation (7) also captures the dimension of this relationship. The first three canonical correlations for the year 1994-2018 are given in Figure 4. The first one varies from 0.7 to 0.83 with no discernible patterns on any years, even in 2008-2009 when the commonality index based on returns was significantly above all other years. Although there is slightly more variation in the second and third canonical correlations, the fluctuation cannot be readily associated with any economic events, and the variations are mild, suggesting a stable relationship between returns and order flows. Note that the regression model in Equation (7) is expected to capture the variation in returns $\mathbf{r_t}$ through the variation in order flows $\mathbf{x_t}$, the rationale being that information is likely to be absorbed in order flows first and this in turn will lead to the variations in returns.

In fact, Hasbrouck and Seppi (2001) postulate that the commonality in order flows is the source of commonality in returns. In our analysis, the largest canonical correlation between returns and order flows for most of the years is close to unity. This indicates that these two variables could be moving together instead. Although the concept of co-integration generally applies to non-stationary series, it can be extended to the regression model in Equation (7) as well. If we write the model (7) in the following error-correction form

$$\mathbf{r_t} - \mathbf{x_t} = (\mathbf{C} - \mathbf{I})\mathbf{x_t} + \epsilon_\mathbf{t}, \tag{11}$$

we can use the smallest canonical correlation between $\mathbf{r_t} - \mathbf{x_t}$ and $\mathbf{x_t}$ to construct linear combination of $\mathbf{r_t} - \mathbf{x_t}$, so that $\mathbf{l'r_t} \sim \mathbf{l'x_t}$. Figure 5 is plotted based on this formulation, and clearly indicates that this is a stable relationship. It is therefore possible that both returns and order flows (and indirectly volume) have common sources of influence[2]. We argue that the high canonical correlations observed for over two decades might also indicate that there may be exogenous economic or behavioral (sentiment or attention) factors driving both returns and order flows.

4.4. Interpretation of canonical components

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 check if there is any economic interpretation of the coefficients of the principal components or the components of the matrices **A** and **B**. After all, they are directly based on returns and order flows. As our results have demonstrated, the first PC's loadings of returns are roughly equal and they reflect the composition of the market factor. By simply examining these coefficients, we do not see that larger firms get more weight consistently. As the constituent stocks

[2] This hypothesis has been raised as early as Tauchen and Pitts (1983).

might vary over time in the Dow Jones index, we check on the consistency via the cosine measure proposed by Krzanowski (1979)[3].

Using this measure and taking the average correlation over the years, we obtain a similarity measure for the returns, of which the results for the first principal component is plotted in Figure 6. As expected, there is a close resemblance between the first PC of returns and the first column of the matrix $\mathbf{A}$ in Equation (7), though the resemblance between the first PC of order flows and the first row of the matrix $\mathbf{B}$ is marginally lower.

Next, we explore here a complementary side of the model in Equation (7), which relates to the topic of portfolio management. The rank condition on the matrix $\mathbf{C}$ also implies that there are linear constraints of type $\mathbf{l'C} = 0$, which results in $\mathbf{l'r_t} \sim \mathbf{l'a_t}$. These $\mathbf{l}$'s also correspond to producing minimum variance for linear combination of $\mathbf{r_t}$. In other words, they can be readily related to mean-variance optimization. If we let $\mathbf{w}$ be the weight vector of the efficient portfolio, it can be shown that this vector belongs to the subspace of PCs that are associated with smallest eigenvalues of the return covariance matrix. To see this, note that if we define $\tilde{\mathbf{A}} = [\mathbf{A}_4, \mathbf{A}_5, ...\mathbf{A}_{30}]$, where $\mathbf{A}_i$ is the i^{th} column of $\mathbf{A}$, then the projection weight vector $\mathbf{w}$ onto $\tilde{\mathbf{A}}$ is $\hat{\mathbf{w}} = \mathbf{P}_{\tilde{\mathbf{A}}}\mathbf{w}$, where $\mathbf{P}_{\tilde{\mathbf{A}}}$ is the projection matrix, that is, $\mathbf{P}_{\tilde{\mathbf{A}}} = \tilde{\mathbf{A}}(\tilde{\mathbf{A}}'\tilde{\mathbf{A}})^{-1}\tilde{\mathbf{A}}'$. Therefore, the cosine angle between $\mathbf{w}$ and $\hat{\mathbf{w}}$ is $\frac{<\mathbf{w},\hat{\mathbf{w}}>}{\|\mathbf{w}\|\|\hat{\mathbf{w}}\|}$. The cosine angle for all years can be close to one. Thus we have demonstrated that the significant canonical components indicate high commonality, while the insignificant ones indicate the diversification aspect.

The loadings on eigenvectors associated with principal components and canonical correlations vary over time. This is due to the changes in both the market conditions and the composition of the thirty Dow constituent stocks that can vary from year to year. For our analysis, we pool the loadings so that we can relate them cross-sectionally to factors that may affect trading patterns. These include price level (measured by the natural logarithm of the stock price) and size (measured by market capitalization), book value (measured by book-to-market ratio), among others. As the sample stocks vary over time, in order to account for the frequency of their inclusion in the Dow Jones index, we conduct a weighted regression model with frequency as weights, and the results are presented in Table 8 Panel A. All columns here are response variables in the regressions.

[3] 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 $\mathbf{S} = \mathbf{L'MM'L}$. If $k = 1$, λ_1 is simply the square of inner product of the two vectors.

The “range” column provides information about the variation of the characteristics used for the regression. The first three columns (Ret PC 1, Order Flow PC 1, and B-coeff) are the first set of regression on price level, market capitalization, and book-to-market ratio. Our results suggest that the variation over the loadings are better explained by the market capitalization and the book-to-market ratio, with both of them indicating that larger stocks with high liquidity tend to dominate the commonality coefficients.

In a separate study (under a different context), where the focus is on market micro-structure behavior and the efficiency of the market, Madhavan (2000) reviews a model proposed by Demsetz (1968) to study the variation in the bid-ask spread. In this study, in addition to market capitalization, they also consider the inverse of the price, trading volume and the volatility of the returns. We regress the PCA loadings on the corresponding firms’ characteristics, the objective being to show that the loadings may be related to some firms’ characteristics or related to its bid-ask spread. However, as the average bid-ask spreads for all 30 stocks in recent year are close to 0.01, we follow the model proposed by Demsetz (1968) to check the relationship between the loading and bid-ask spread. The weighted regression is presented in Table 8 Panel B. Interestingly, the price coefficient is not significant in any of the regressions. This may be due to the restrictive range of prices included in the top thirty stocks. Another observation worth noting is that the first component of return is a function of both market capitalization and volatility, but the second component of returns and order flows are influenced by volatility in return. Highly volatile stocks tend to have higher coefficients.

4.4.1. Relationship with 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 carry a certain similarity, and there is some commonality among their behavior. Typically the co-movement models are VAR models, such as $\mathbf{r_t} = \mathbf{\Phi r_{t-1}} + \mathbf{e_t}$ (not regression models of type (7)), and an appropriate linear combination of r_t will be free of this co-movement. The VAR(1) model on the returns has been considered by various authors. Among these work, DeMiguel, Nogales, and Uppal (2014) study the serial dependence of returns and show empirically that the mean-variance portfolios based on the VAR model outperform traditional portfolios. We want to explore whether the model in Equation (7) can be improved by accounting for the serial correlation in the residuals. This can be incorporated via two different models:

$$\begin{aligned} &\mathbf{r_t} = \mathbf{C}\mathbf{x_t} + \mathbf{u_t} \quad \text{where} \qquad \mathbf{u_t} = \phi\mathbf{u_{t-1}} + \mathbf{a_t}, \\ &\quad \text{or} \\ &\mathbf{r_t} = \mathbf{\Phi}\mathbf{r_{t-1}} + \mathbf{C}\mathbf{x_t} + \mathbf{a_t}. \end{aligned} \tag{12}$$

We choose the first model as our focus is on the influence of the order flow on returns.

Results of the index (10) based on the elements of the matrix **C** are presented in Figure 7. Our analysis reveals that there is not much difference between the two models. This is potentially due to the fact that co-movement refers to time series behavior and commonality refers to cross-sectional behavior. In any case, the higher values in 2008 and 2009 confirm the dependence of stock behavior on market volatility.

5. Sources of commonality

Our hypothesis is that investors' sentiment and attention are the sources of commonality. Here we move on to discuss how these sources generate information that may lead to commonality or co-movement of stock returns, order flows, and liquidity.

In existing literature, the consensus is that the most prominent factors of commonality are 1) information diffusion, and 2) common stock holding by investors. However, it is not possible to know the exact composition of every individual investor's portfolio. Moreover, it is the trading activity itself that generates commonality. Since 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 immediately. We focus here on the use of twitter data in extracting stock related information via appropriate natural language processing methods. The analysis of investor sentiment data to provide market signals has been a focus in recent research. In this section, we illustrate how the PCAs of news analytics (commonality indices) are related to the well known sentiment index introduced by Baker and Wurgler (2006), as well as the attention measure in Chen et al. (2019). We will follow up with sentiment and attention data that is collected or computed in real time, and elaborate on how it can help to explain the commonality in both returns and order flows. This serves as a strong confirmation that the discussion in news media is an important source of commonality. A major advantage of our proposed approach by gauging sentiment and attention based on news analytics is that this relationship can be exploited for developing real time trading strategies as well.

To construct an index measure from the information provided, we consider the sentiment measure variables presented in Table 10 Panel A along with the attention measure variables list in Table 10 Panel B.

5.1. Low Frequency Analysis: Monthly Sentiment and Attention

To relate the commonality indices to sentiment and attention, we first extract the principal component loadings for each year for both returns and order flows, and then construct PCA score for each month using the monthly averages of returns and order flows. To achieve this, we generate the principal components (eigenvectors) and the **A** and **B** matrices from the annual data, and then multiply the monthly average of returns and order flows to get the PCA score for each month[4]. The time series plots of these PCA scores and the sentiment index of Baker and Wurgler (2006) are presented in Figure 8. We also include the **A** and **B** components in the regression model as they provide some significant results. It is clear that order flow commonality exhibits higher variance in recent times, though both returns and order flow are highly volatile in 2008. The autocorrelation in order flow indicates that lag 1 is positive at approximately 0.225. The sentiment score and the **B** score are non-stationary with partial autocorrelations approaching 1. In addition, the **A** score is stationary but attention index is non-stationary with partial autocorrelations greater than 0.8. From the plots, it appears that these series are likely to be closely related to sentiment and attention indices. Regression results given in Table 9 further confirm the findings. Here, PC_Ret is the monthly PC return scores, "Border" and "A-invRet" are the monthly number. In Equation (7), if both side take the inverse of A, it becomes $\mathbf{A^{-1}r_t} = \mathbf{Bx_t} + \mathbf{A^{-1}a_t}$. The columns are response variables, and provide information about the optimal mix. For instance, the 3^{rd} column (**B** order) is only regressed against "Sentiment" with a constant, and the R^2 is 27.86%. But if we add the

[4] Alternatively we could have constructed these using PCA loadings based on monthly data, but they may not provide stable behavior due to the smaller sample size.

"Attention" to the regression (the 4^{th} **B** order column), then we can achieve an R^2 of 31.33%.

From Table 9, we can see that although the order flows' influence on returns is small, it is still significant. However, sentiment scores do not appear to be significant here. This is potentially due to the monthly aggregation of these scores for returns and order flows, though it was done out of necessity to match with sentiment scores, available only on a monthly level. It is also worth noting that most of the variables in the Baker and Wurgler (2006) index are slow-moving, but returns and order flows are fast changing, and their inter-relationship can be better captured with the high-frequency sentiment data.

Note that scores that result from the **B** component in the regression model is highly correlated with the sentiment index, and the scores from the **A** component is highly correlated with attention index. These coefficients are not obtained from maximizing the variance of linear combination as in PCA, but rather from maximizing the correlation between the order flows and the returns. In other words, canonical correlation analysis results in the optimal combination that maximises the correlation. From the regression results in Table 9, it is obvious even after accounting for the commonality values from PCA of order flows, sentiment scores play an important role. By appropriately rescaling the **B** coefficients that in most normal periods when there is no volatility in sentiments, there is a closer alignment between these coefficients and the sentiment scores, as can be seen in Figure 9. Similarly, the **A** component score is highly related to attention measure.

5.2. High Frequency Analysis: Intraday Sentiment and Attention

The analysis performed in the previous section relies on low frequency sentiment and aggregated attention data, and hence its influence on returns commonality is not easy to detect. In this section, we move on to discuss the use of high frequency sentiment data. There has been a recent surge of academic interest in this topic (see Mitra and Mitra (2011) and Peterson (2016)).

In order to extend the results observed in a low frequency context in the previous section, in particular the notion that the commonality scores (that result from return-order flow regression relationship) on returns are associated with 'attention', while scores on the order flow are associated with 'sentiment', we need to develop some high frequency 'attention' measures. We use some of the measures mentioned in Chen et al. (2019). These can be taken as attention proxies at the stock level, and are tabulated in Table 10 Panel B.

Before we describe the analysis, a general comment is in order. Most of the published research on investor sentiment and attention focus on using the sentiment or attention data for predicting stock returns. However, our goal here is to use this data to study the commonality and co-movement of stocks. To achieve this, we first perform multivariate regressions, where the predictors include each stock's own sentiment and attention proxies, and as well as other stocks' proxies. Our results indicate that the dominant factor is the order flow variables—including the stock's own order flow as well as the order flows of all other stocks. To illustrate this, we calculate the R^2 values that result from (1) each stock's return from its own order flow, (2) each stock's return from its own order flow and an index constructed as equal-weighted sentiment and attention proxies, and (3) with other stocks information added to (2). For the sentiment and attention measures, we have tested indices based on principal components as well as equal-weighted scores of normalized proxies. We find that the equal-weighted index gives better results. The summary graph is presented in Figure 10[5]. The higher R^2 values for the model in Equation (3) clearly demonstrates the presence of substantial commonality in all the variables among the stocks.

Similar to the monthly analysis in the previous section, we can study commonality at the high frequency level using the indices that result from PCA and CCA. To better understand the dichotomous relationship between returns via-à-vis attention and order flows via-à-vis sentiment, we want to see how the two dimensions (sentiment and attention) are related among themselves. Chen et al. (2019) mentioned that the twelve attention measures they construct do clearly capture a dimension different from the sentiment dimension. The attention measures we consider here are limited to five (a subset of their full list) due to the high frequency nature of our analysis. These attention measures are related to the past market performance of the stocks, whereas the sentiment measures are based on voluntary feedback from the investors. It is therefore possible that the investors' feedback could depend upon the past performance of the stocks themselves.

We examine the interdependence of the attention and sentiment dimensions via factor analysis of the ten proxies—five from each dimension. The factor analysis is essentially based on PCA, but augmented by rotation of the loadings according to a predetermined criterion. Here we use the popular varimax rotation[6]of first two factors, as they account for almost 45% of the total variance.

[5] For the years 2008 and 2009, the R^2 values do not include sentiment data as they were not available, and so the numbers in Model (2) are based on only order flow and attention index.

Results are given in Table 11. The first factor is predominantly loaded on the attention proxies, while the second factor is mainly loaded on the sentiment proxies. Among the two, factors loadings are on the buzz and the tweet volume ratio on the sentiment side, and the abnormal trading volume and the net flow on the attention side. To a large extent, these are intensity measures along with the net flow that may indicate market imbalance. This analysis is further supported by the canonical correlation values, which range from 0.37 to 0.49 over the years. In theory, if these dimensions are independent, then the correlation should be zero. If they are virtually identical, then we should expect it to be close to one. The magnitude of these correlations indicate that there is some overlap among the two dimensions.

Next, we want to relate the commonality scores at the high frequency level to attention and sentiment. Given the presence of noise due to market friction and other factors in the high frequency data, the overall R^2 values are not expected to be as high as in Table 9 for the monthly data. Nevertheless the relationships obtained in our analysis at the high frequency intraday level are broadly consistent with the monthly analysis in the previous section. To select a single component measure for attention and sentiment dimensions, first we run PCA over 30 stocks on each measure, and then run PCA again on the first component scores. Table 12 presents the commonality, attention, and sentiment relationship at the high-frequency intraday level. All columns presented are response variables, and regressions are run against different combinations of sentiment and attention metrics to determine which combination yields the better R^2. It can be inferred from these results that at the high frequency level, both sentiment and attention scores play a significant role, though it is not possible to judge conclusively if the attention and sentiment dimensions are as uniquely aligned with either returns or order flows. Our results can be improved further if intraday data for the remaining measures can be constructed and used for the analysis.

Note also that the change in the sentiment score and the recent cumulative returns have positive impact on the commonality indices resulting from the regression model. Moreover, the sAS proxy (nearness to monthly high) has a positive impact on order flow index, while the change in the recent sentiment has a positive impact on the return index. These influences of attention and sentiment variables can be expected from a behavioral point of view. Commonality also exhibits some intraday effects, where higher activities during market opens and closes lead to higher commonality (see

[6] Varimax is an orthogonal rotation that minimizes the number of variables that have high loadings on each factor.

Figure 11).

5.3. Related Work on Sources of Commonality

The literature on the study of commonality generally attribute its source to the common variation in the demand or supply of liquidity. Recent studies have attributed the source of commonality to groups that have come to dominate trading. Harford and Kaul (2005) argue that a stock being a constituent stock of an index such as S&P500 can result in co-movement of stocks, a feature not found in a sample of non-index stocks. The differences are exhibited in the percentage of variation explained by the first factor, as well as the correlation between first component and the aggregated S&P being nearly perfect. Here no distinction is made about the demand and supply sides. It fits into how investors categorize the stocks and focus on the categories of their interest.

Green and Hwang (2009) use the price level of a stock to demonstrate that similarly priced stocks move together because investors categorize stocks according to nominal price level. Stock splits that result in lowering nominal prices provide a natural way to test this hypothesis. Generally, institutional investors hold high-price stocks, while retail investors hold low-price stocks. Splits are more attractive to retail investors due to their wealth constraints. Split stocks experience a shift in the co-movement with an increase in low-price stocks and a decrease in high-price stocks. In this explanation, the shift is not explained by the firms' fundamentals, but rather merely by investor sentiments.

Corwin and Lipson (2011) suggest that the commonality is driven by professional traders through their program trades, and these trades are likely to be highly correlated; the correlation may be due to similar reading of common signals. Koch et al. (2016) attribute commonality to how mutual funds are traded. After estimating a stock's commonality, it is linked to mutual fund ownership. Malceniece et al. (2019) suggest that the increase in liquidity co-movement is due to an increased presence of high frequency traders in the market.

The relationship between investor sentiment and stock returns has been actively explored in finance. With the real time dissemination of electronic data, research in the high frequency setting is now possible. Kumar and Lee (2006) show how the retail investors sentiment 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. The retail investors' trades tend to be correlated, and therefore buy-sell imbalance across non-overlapping portfolios is taken as an indicator of investors sentiment. The cross-sectional regression model includes various control factors, including Fama-French and momentum factors. It is shown that using lead-lag relationships, the sentiment measures exert impact on small stocks, stocks with lower nominal prices, and 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) 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 left out of the index in recent versions. The monthly index correlates with major stock market movements. The review paper by Zhou (2018) outlines other measures as well. Here we want to first examine at the low frequency (monthly) level, how the commonality indices on returns and order flows correlate with the Baker and Wurgler's sentiment index.

The role of media coverage that disseminates information to a broad audience has been extensively elaborated. Fang and Peress (2009) observe that stocks with high media coverage tend to earn less than stocks with no media coverage based on a cross-sectional study, which holds true even after controlling for the widely-recognized risk factors. At the macro level, how news articles can predict the returns and volatilities etc has been well documented by Calomiris and Mamaysky (2019). But to extend this into the study of co-movement at the high frequency level, we need to look for how sentiment is formed instantaneously through the news media and how it is impacting the stock performance. Recent research also focuses on how managers who have access to better information form sentiment and suggest measures that lead to their attention (see Jiang, Lee, Martin, and Zhou (2019)). It is argued that managers sentiment can be different from investors sentiment and they provide a different dimension to behavioral aspect of investing. Here we want to unify the two related streams of work by parsing through the variables that make up these measures, and how they influence the return and order flow commonalities differently.

Jiang et al. (2019) summarize various sentiment measures starting with the consumer confidence indices released by University of Michigan and Conference Board that have been long considered to be good barometers of future economic outlook. They also suggest fourteen monthly economic variables that may be aligned with macroeconomic business cycles. Two predictors that seem to stand out are stock return volatility and the long-term government bond yield. One would expect that the sources of commonality could arise from predictors that are significant rather than insignificant ones. While many of these variables are not at the stock level, they do extract the managers sentiment from the textual tone of the conference calls on financial reporting of the company performance.

The attention proxies have been widely noted in the literature. Chen et al. (2019) list them as follows: Abnormal trading volume, calculated as the ratio of trading volume to the average over the previous year; Extreme returns, arrived at a similar manner; Past monthly cumulative return over the prior year; Analyst coverage of earnings per share forecasts; Changes in the advertising expenses. While these are at the stock level and are aggregated to the market level, a market indicator for nearness of prior year NYSE index to historical high is also suggested. An index is arrived at either by equal weighing of these measures, or by PCA or by Partial Least Squares methods. Some of the attention measures, as they are available on a continual basis can be used in a high-frequency context as well. We will return to this later. Because of our focus is on co-movement and commonality, the effect of attention or inattention can lead to stock return co-movement. This is investigated by Huang, Huang, and Lin (2019) who demonstrate that when there are external events then can grab the investors attention, they try to focus more on learning about market shocks than about firm-specific shocks.

6. Conclusions

We have performed a comprehensive study of commonality using an ensemble of modeling framework including reduced-rank regressions, canonical correlation analyses, and principal component analyses. Our longitudinal study established that commonality in returns, order flows, and liquidity is persistent, and holds stable for an extended period of time. We then move on to investigate why commonality has persisted over time. The majority of the literature on the study of commonality attributes the commonality on return to the commonality on order flows, and employ "demand-side" or "supply-side" explanation to justify the commonality on order flows. In this paper, we model both returns and order flows endogenously, and argue that their commonality are driven by exogenous factors, notably investors' sentiment and attention. We leverage our analysis on a recent stream of research focusing on behavioral finance. We provide empirical evidence showing that commonality in returns are driven by investors' attention, while commonality in order flows are caused by investors' sentiment, thus demonstrating the multi-dimensional aspect of commonality. Our results are robust and hold true when we perform our analyses in both low-frequency and high-frequency context. These analyses not only extend the knowledge of commonality in asset characteristics, but also open up new possibilities to optimize investment or trading strategies.

REFERENCES

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.

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., 2019. Investor attention and stock returns. Working paper .

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

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.

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.

Demsetz, H., 1968. The cost of transacting. The Quarterly Journal of Economics 82, 33–53.

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.

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

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.

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.

Jolliffe, I., 2002. Principal Component Analysis, second edition. Springer-Verlag, New York.

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.

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

Madhavan, A., 2000. Market microstructure: A survey. Journal of Financial Markets 3, 205–258.

Malceniece, L., Malcenieks, K., Putniņš, T. J., 2019. High frequency trading and comovement in financial markets. Journal of Financial Economics, *Forthcoming* .

Mitra, G., Mitra, L., 2011. The handbook of news analytics in finance, vol. 596. John Wiley & Sons.

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

Peterson, R. L., 2016. Trading on sentiment: The power of minds over markets. John Wiley & Sons.

Reinsel, G. C., Velu, R., 1998. Multivariate reduced-rank regression: theory and applications, vol. 136. Springer Science & Business Media.

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.

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.

Table 1

Definition of order flow and liquidity measures.

Variables	Definition
Return	$\log\left(\frac{P_{A_t}+P_{B_t}}{2}\right)-\log\left(\frac{P_{A_0}+P_{B_0}}{2}\right)$
Order Flow Measures:	
Total number of trades	n_t
Total share volume	$\sum_{j=1}^{n_t} v_j$
Total dollar volume	$\sum_{j=1}^{n_t} \log(P_j)v_j$
Square root of dollar volume	$\sum_{j=1}^{n_t} \sqrt{\log(P_j)v_j}$
Signed trades	$\sum_{j=1}^{n_t} \text{sign}(v_j)$
Signed share volume	$\sum_{j=1}^{n_t} \text{sign}(v_j)v_j$
Signed dollar volume	$\sum_{j=1}^{n_t} \text{sign}(v_j)\log(P_j)v_j$
Signed square root dollar volume	$\sum_{j=1}^{n_t} \text{sign}(v_j)\sqrt{\log(P_j)v_j}$
Liquidity Measures:	
Quoted spread	$P_A - P_B$
Proportional quoted spread	$(P_A - P_B)/P_M$
Depth	$\frac{1}{2}(Q_A + Q_B)$
Effective spread	$2\left\|P_t - P_M\right\|$
Proportional effective spread	$2\left\|P_t - P_M\right\|/P_t$
Amihud illiquidity measure	$\|r_t\|$/Volume

P_j, v_j and sign(v_j) are the price per share, share volume and the direction of j-th trade, respectively. For liquidity measures, P denotes price and subscripts indicate: t=actual transaction, A=ask, B=bid, M=bid-ask midpoint. Q denotes the quantity guaranteed available for trade at the quotes, (with subscripts: A=ask, B=bid). All 15-min aggregated liquidity measures except Amihud illiquidity measures are time-weighted, according to the number of seconds that the quote prevailed. We generate Amihub illiquidity measure where r_t is 15-min log midpoint return and Volume is 15-min dollar trading volume.

Table 2
Providers of web analytics

Provider	Description
Ravenpack (NewsScore)	The key information on entities (company, organization, currency, commodities and place) from major news sources are gathered and processed for their relevance and are assigned event sentiment score. The scores are aggregated on a rolling window basis.
Thomson Reuters (News Analytics)	Data fields include relevant score of the news item to the asset, number of sentiment words or tokens, sentiment classification, novelty of the content, feed volume etc. The sources are Reuters, PR Newswire etc.
Thomson Reuters (MarketPsych Indices)	News and social media information in real time are delivered as data series; there are categorized into three types of indicators: Emotional indicators, Macroeconomic metrics and Buzz metrics on the asset level. The data is updated on a minute basis. The social media data includes blogs, internet-forums and finance-specific tweets.
iSentium (iSense)	Extracts sentiment signals using natural language processing architecture on Twitter data; the data feed includes retweets, if the author has a finance-related bio, number of followers etc. The impact is measured by the number of retweets for each authors tweets.
Social Market Analytics (SMA)	The data set provides a snapshot of sentiment factors at a 15 minute interval sourced from Twitter/StockTwists messages. The raw scores are adjusted for the average and volatility. A measure of unusual volume activity is also provided.

Table 3

Descriptive statistics of select variables from Table 1

Year	Return (basis point)	Std Dev. of daily return *100(%)	S&P 500 return (basis point)	Std Dev. of S&P 500 return *100(%)	Signed Trades	Turnover *100(%)	Amihud illiquidity measure (10ˆ-10)	Proportional quote spread *100(%)	Investor Sentiment
1994	3.22	1.37	-0.42	0.62	36	0.24	1.90	0.38	0.66
1995	14.39	1.37	11.77	0.49	73	0.27	1.40	0.30	0.22
1996	9.65	1.61	7.54	0.74	112	0.27	1.25	0.28	0.66
1997	14.82	1.94	11.33	1.14	195	0.33	1.02	0.21	0.75
1998	14.20	2.27	10.20	1.28	219	0.32	0.81	0.17	0.19
1999	6.49	2.39	7.72	1.14	446	0.35	0.61	0.16	-0.02
2000	-8.04	3.59	-3.27	1.40	508	0.64	0.44	0.17	1.12
2001	-3.83	2.72	-4.72	1.36	386	0.45	0.45	0.10	2.23
2002	-6.85	2.58	-9.22	1.64	193	0.41	0.50	0.09	0.16
2003	9.49	1.64	9.87	1.08	355	0.40	0.35	0.05	-0.52
2004	2.32	1.24	3.66	0.70	245	0.36	0.25	0.04	-0.07
2005	-0.07	1.16	1.38	0.65	-55	0.45	0.19	0.03	0.22
2006	6.52	1.16	5.29	0.63	213	0.57	0.16	0.03	0.37
2007	3.03	1.40	1.89	1.01	-628	0.62	0.13	0.03	0.56
2008	-12.66	3.53	-15.87	2.58	-206	1.24	0.25	0.04	-0.09
2009	11.83	2.37	9.83	1.72	-616	1.01	0.18	0.03	-0.65
2010	4.68	1.41	5.42	1.14	-259	0.83	0.12	0.03	-0.59
2011	2.15	1.77	1.07	1.47	-384	0.81	0.14	0.03	0.16
2012	6.03	1.21	5.36	0.80	-369	0.62	0.12	0.03	-0.08
2013	9.95	1.15	10.54	0.70	-120	0.55	0.11	0.03	0.07
2014	4.61	1.15	4.54	0.72	-15	0.49	0.10	0.03	-0.02
2015	3.47	1.40	0.19	0.98	-122	0.50	0.11	0.03	0.02
2016	5.15	1.31	3.95	0.82	-114	0.50	0.10	0.02	-0.17
2017	7.81	0.97	7.16	0.42	-256	0.41	0.08	0.02	-0.15
2018	0.14	1.59	-1.99	1.07	-833	0.51	0.10	0.03	-0.08

The sample is the top 30 largest stocks sorted annually by market capitalization(size). The sample period extends from January 1994 to December 2018. The annual investor sentiment is the simple average of monthly sentiment got from Jeffery Wurgler's website. Other measures are the daily averages across 30 stocks.

Table 4

Principal Component Coefficients for eight order flow measures

	1994-2006		2007-2018	
	First	Second	First	Second
Number of Trades	0.37	-0.30	0.50	0.03
Sum dollar volume	0.40	-0.31	0.50	0.03
Sum share volume	0.40	-0.31	0.50	0.03
Sum sqrt dollar volume	0.41	-0.32	0.50	0.03
Signed Trade	0.30	0.33	-0.03	0.48
Signed dollar volume	0.31	0.41	-0.03	0.50
Signed share volume	0.31	0.41	-0.03	0.50
Sigend sqrt dollar volume	0.33	0.40	-0.03	0.51
Eigenvalues	4.02	2.87	3.87	3.24

Table 5

Principal Component Coefficients for six liquidity measures

	First	Second	Third
Dollar Depth	-0.29	0.64	0.00
Shares Depth	-0.27	0.66	0.01
Effective spread	0.47	0.17	-0.54
Prop. effective spread	0.48	0.24	-0.42
Quoted spread	0.44	0.14	0.54
Prop. quoted spread	0.45	0.23	0.49
Eigenvalue	3.06	1.77	0.68

Dollar Depth is the simple average of the best bid quote and best ask quote in dollars. Shares Depth is the similar definition in shares. Sample period covers from January 1994 to December 2018.

Table 6

Patterns of commonality in return, order flow and liquidity

Variables	Years	Rest of the day eigenvalue with CI	Last hour eigenvalue
Return	All	8.84±0.03	12.75
	2008	13.18±0.25	19.34
Signed Trade	All	4.56±0.02	5.68
	2008	7.41±0.14	8.69
Amihud's illiquidity	All	1.47±0.01	1.70
	2008	9.61±0.19	12.30

The table presents the largest eigenvalue from the principal component analysis for different variables, for different times in the day and for the two different sample periods. All variables are calculated at 15-minute intervals.

Table 7

Principal Component analysis for return, order flow and liquidity measures

	Return				Signed trade				Amihud's illiquidity			
Year	Eigenvalues			Cum. explained	Eigenvalues			Cum. explained	Eigenvalues			Cum. explained
	1	2	3	variance(%)	1	2	3	variance(%)	1	2	3	variance(%)
1994	6.22	1.14	1.06	28.06	3.59	1.59	1.23	21.39	1.68	1.32	1.15	13.83
1995	6.06	1.34	1.16	28.53	3.48	1.35	1.30	20.44	1.46	1.18	1.17	12.68
1996	8.60	1.22	1.06	36.28	4.71	1.61	1.38	25.69	2.09	1.27	1.13	14.96
1997	11.23	1.17	1.00	44.63	6.66	1.45	1.18	30.99	2.97	1.21	1.16	17.80
1998	9.81	1.20	1.09	40.33	6.01	1.87	1.21	30.32	3.10	1.30	1.20	18.65
1999	8.00	1.52	1.15	35.58	5.75	1.49	1.30	28.44	2.53	1.26	1.16	16.49
2000	6.68	2.17	1.22	33.55	4.69	1.95	1.42	26.88	2.25	1.61	1.17	16.78
2001	7.58	2.27	1.32	37.25	6.69	1.84	1.34	32.90	3.51	1.46	1.13	20.33
2002	10.62	1.41	1.18	44.06	9.04	1.27	1.12	38.12	5.75	1.24	1.16	27.16
2003	12.06	1.30	1.03	47.99	8.55	1.19	1.02	35.88	5.98	1.48	1.12	28.56
2004	9.30	1.34	1.18	39.42	8.57	1.23	1.15	36.51	4.31	1.33	1.21	22.83
2005	8.99	1.50	1.26	39.14	8.18	1.37	1.19	35.82	4.14	1.38	1.30	22.73
2006	7.65	1.96	1.24	36.16	6.45	1.48	1.29	30.72	3.96	1.65	1.32	23.08
2007	10.56	1.89	1.44	46.31	3.63	1.97	1.26	22.89	4.88	1.53	1.32	25.81
2008	14.33	1.99	1.37	58.96	7.70	1.64	1.36	35.65	10.00	1.51	1.38	42.97
2009	12.49	2.01	1.31	52.71	7.64	1.64	1.22	35.02	7.42	1.46	1.42	34.30
2010	12.50	1.82	1.22	51.82	6.08	1.49	1.11	28.92	6.86	1.50	1.31	32.22
2011	13.46	1.66	1.28	54.66	5.08	1.40	1.24	25.72	7.62	1.47	1.45	35.14
2012	10.04	2.07	1.39	44.99	4.27	1.50	1.25	23.38	5.35	1.62	1.31	27.62
2013	9.31	1.93	1.35	41.92	4.08	1.41	1.22	22.38	4.85	1.57	1.29	25.70
2014	9.48	1.91	1.46	42.80	4.83	1.61	1.32	25.86	5.15	1.42	1.29	26.19
2015	12.57	1.67	1.46	52.31	5.72	1.40	1.28	28.02	7.08	1.51	1.45	33.43
2016	10.78	2.61	1.65	50.10	5.11	1.64	1.40	27.20	6.57	1.78	1.48	32.77
2017	6.01	2.82	2.53	37.86	3.25	1.78	1.57	21.99	4.33	1.79	1.64	25.87
2018	12.86	3.00	1.68	58.47	4.23	1.69	1.30	24.03	7.79	1.94	1.54	37.57

The table lists the first three eigenvalues from the principal components analysis based on standardized data along with the proportion of total variance explained by these three eigenvalues.All variables are calculated at 15-minute intervals.

Table 8

Stock Characteristics and PCA/CC loadings

Panel A

Characteristics	Range: Min-Med-Max	PC1($\mathbf{r_t}$)	PC1($\mathbf{x_t}$)	**B** coeff
log(Price)	[1.40-1.81-2.88]	0.002	0.006	-0.000010
log(Cap)	[7.18-7.97-8.60]	0.034***	0.047***	0.000029***
BM ratio	[0.019-0.270-1.005]	0.022***	-0.005	-0.000070
Constant		-0.108	-0.213***	-0.000220***
R^2		23.58%	25.07%	19.89%

***,**,* denote statistical significance at the 1%,5% and 10% level, respectively.

Panel B

Characteristics	Range: Min-Med-Max	PC1($\mathbf{r_t}$)	PC2($\mathbf{r_t}$)	PC1($\mathbf{x_t}$)	PC2($\mathbf{x_t}$)	**B** coeff
log(Cap)	[7.18-7.97-8.60]	-0.028*	-0.02	0.053***	0.07	-0.000004
Inv Price	[0.013-0.017-0.048]	-0.082	-1.29	-0.051	-0.22	-0.000390
Volatility	[0.010-0.019-0.056]	1.385***	5.86***	1.134**	5.33***	-0.000167
log(Volume)	[13.34-15.66-17.52]	0.003	0.04*	-0.0007	-0.02	0.000012***
Constant		-0.127*	-0.63*	-0.254***	-0.33	-0.000146**
R^2		37.44%	34.51%	32.93%	11.16%	29.04%

***,**,* denote statistical significance at the 1%,5% and 10% level, respectively.

Table 9

Commonality and Sentiment Relationship

	$\mathrm{PC}(\mathbf{r_t})$	$\mathrm{PC}(\mathbf{r_t})$	$\mathbf{Bx_t}$	$\mathbf{Bx_t}$	$\mathbf{A^{-1}r_t}$	$\mathbf{A^{-1}r_t}$
$\mathrm{PC}(\mathbf{x_t})$		0.000001**				
Sentiment	-0.000023	-0.000020	0.001615***	0.001151***		0.028
Attention				0.001894***	8.92**	
Constant			0.001084***	0.001042***	1.95*	0.227***
R^2	0.14%	2.23%	27.86%	31.33%	1.79%	0.51%

"$\mathrm{PC}(\mathbf{r_t})$" and "$\mathrm{PC}(\mathbf{x_t})$" represent the first principle component scores of returns and order flows. Sentiment scores used in the regressions come from Wurgler's website. Aggregated market attention index is obtained via Guofu Zhou's website. "$\mathbf{Bx_t}$" is the product of the row vector "β" estimated from Model 7 annually and the monthly average of order flows vectors. "$\mathbf{A^{-1}r_t}$" is the product of the generalized inverse of vector "α" and the monthly average of returns vectors. The generalized inverse of vector of "α" and "β" are also known as the first canonical correlation coefficients. Sample period is from 1994 January to 2018 December. ***,**,* denote statistical significance at the 1%,5% and 10% level, respectively.

Table 10
Sentiment and attention measures

Variable	Defination
Panel A	
Normalized Sentiment Score (sNSent)	Exponentially time weighted summation of sentiment of unique tweets in the 24-hour window divided by 20-day moving average at the time of observation. Exponential weighting places importance to recent tweets and the 20-day mean provides a baseline.
Tweet Volume Ratio (sNVol)	Ratio of unique tweets in 24-hour interval to the 20-day moving average at the time of observation.
Dispersion (sDisp)	Ratio of unique accounts to volume, measuring the source diversity.
Buzz (sBuzz)	Measure of unusual activity compared to universe of stocks, that provides a cross-sectional view.
Delta (sDelta)	Change in the sentiment score over a 15-min look-back period.
Panel B	
Abnormal Trading Volume (sAVol)	It is well-known that the trading volume exhibits strong intraday pattern. We compute the ratio of trading volume to the average of trading volumes during the same time interval for the previous twenty-two trading days. Or alternatively we could model the intraday patterns through ARMA models or simply by exponential weighted average, that may pick up recent trends.
Past Returns (sAPR)	The cumulative return of previous twenty-six 15-min time intervals.
Extreme Returns (sAER)	Ratio of returns at the end of the 15-min interval to the average of previous twenty-six time intervals as there is no reason to expect any intraday periodicity in returns. The ratio will be pruned of any single outlier observation that may have an undue effect.
Nearness of Monthly High (sAS)	Ratio of current price to the highest price in the last twenty two days.
Inflow / Outflow measures (sNetF)	Ratio of total dollar volume of buyer (seller)-initiated trades in one 15-min time interval to the average of same measures during the same time interval for the previous twenty-two trading days. We take the difference as the net flow measure. This is a proxy for order imbalance.

Table 11

Factor analysis on attention and sentiment proxies

		Factor 1	Factor 2
Attention	sAER	0.00	0.00
	sAPR	-0.22	0.00
	sAS	-0.21	0.07
	sAVol	**0.97**	0.24
	sNetF	**0.97**	0.23
Sentiment	sNSent	-0.03	0.41
	sDisp	-0.03	-0.33
	sBuzz	0.06	**0.75**
	sDelta	0.00	-0.05
	sNVol	0.08	**0.99**

This table presents the varimax results of 5 attention proxies and 5 sentiment proxies.

Table 12

Commonality, attention and sentiment relationship at the high-frequency level

	PC($\mathbf{r_t}$)	PC($\mathbf{x_t}$)	$\mathbf{Bx_t}$	$\mathbf{Bx_t}$	$\mathbf{A^{-1}r_t}$	$\mathbf{A^{-1}r_t}$
Sentiment	0.0001417***	71.08***	0.0003045***		8.075***	
sNSent						-8.21***
sDelta				0.0003072***		16.41***
Attention	0.0002020***	32.73***	0.0002046***		11.557**	
sAPR				0.0003237***		21.07***
sAS				0.0000479***		
Constant			0.0000217	0.0000237	-0.061***	0.05
R^2	1.53%	2.33%	3.02%	1.25%	5.74%	2.88%

We run PCA over thirty stocks on each measure and run PCA again on the first PC scores to obtain the composite sentiment and attention indices. The variables "sNSent","sDelta","sAPR","sAS" are the first principal scores of normalized sentiment score, delta sentiment score, past returns and nearness of monthly high, respectively. The commonality indices on returns and order flows, whether they are PC's or are based on model (7) are computed exactly as in low frequency data, instead using data on 15-min intervals. Sample period is from 2010 January to 2018 December. ***,**,* denote statistical significance at the 1%,5% and 10% level, respectively.

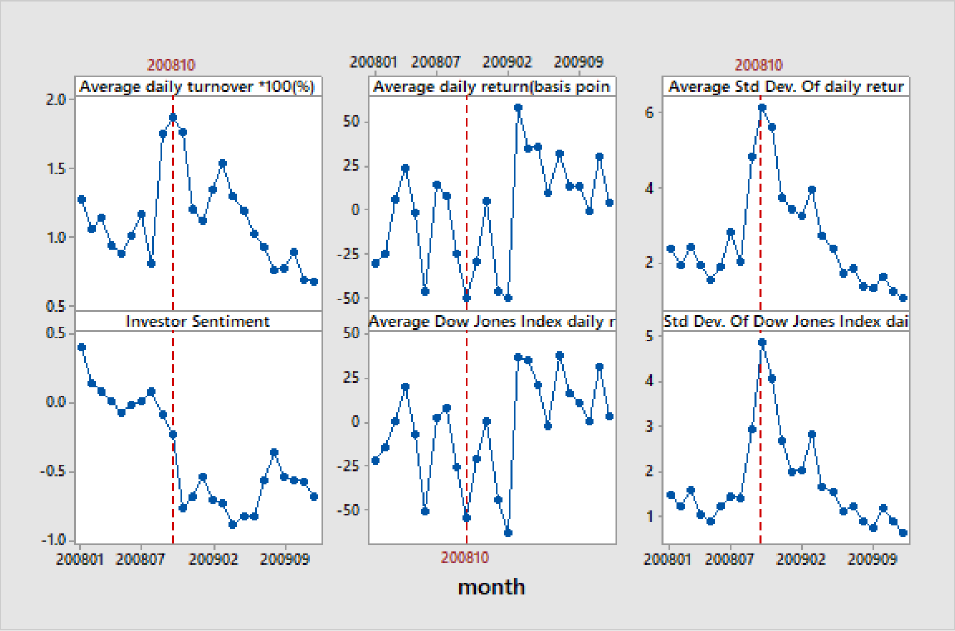

Figure 1. Select variables' performance during months of turbulence (Jan.2008 to Dec.2009)

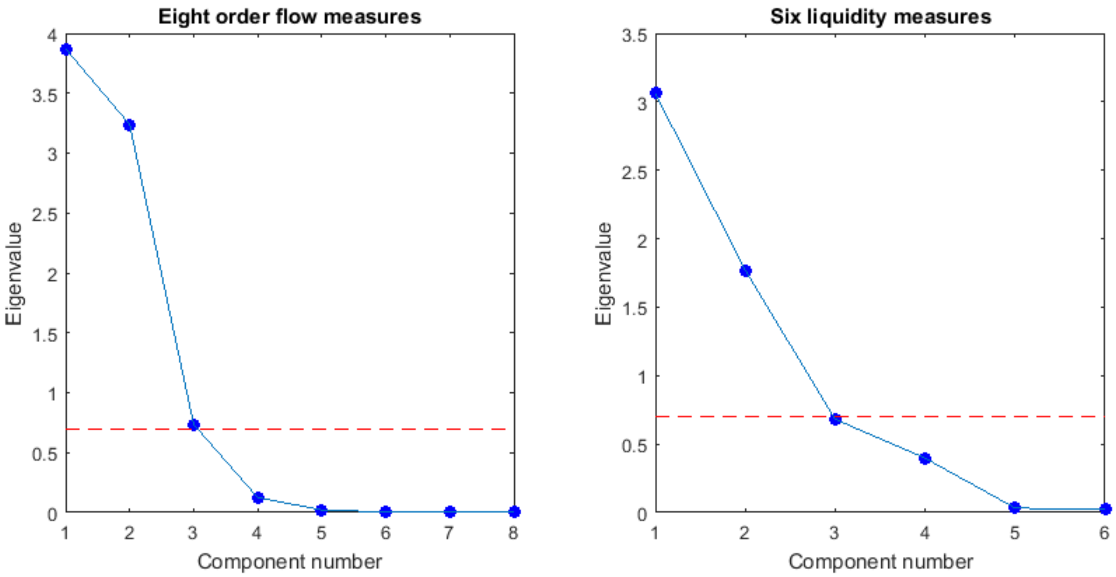

Figure 2. Distribution of eigenvalues for order flow and liquidity measures (Jan.1994 to Dec.2018)

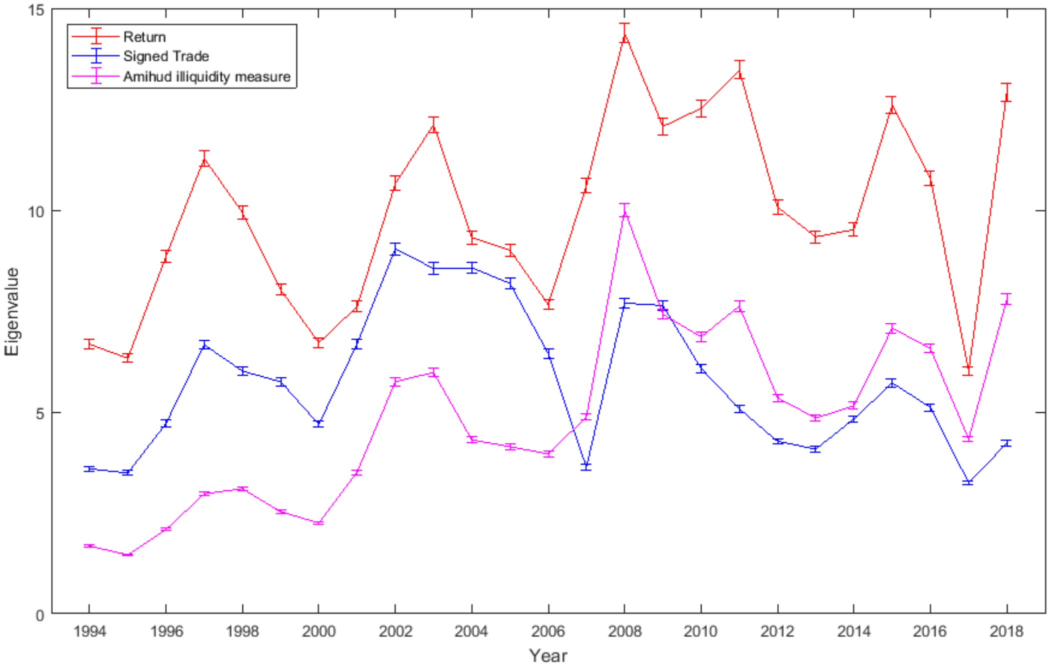

Figure 3. Commonalities in return, order flow and liquidity with 95% CI over time (Jan.1994 to Dec.2018)

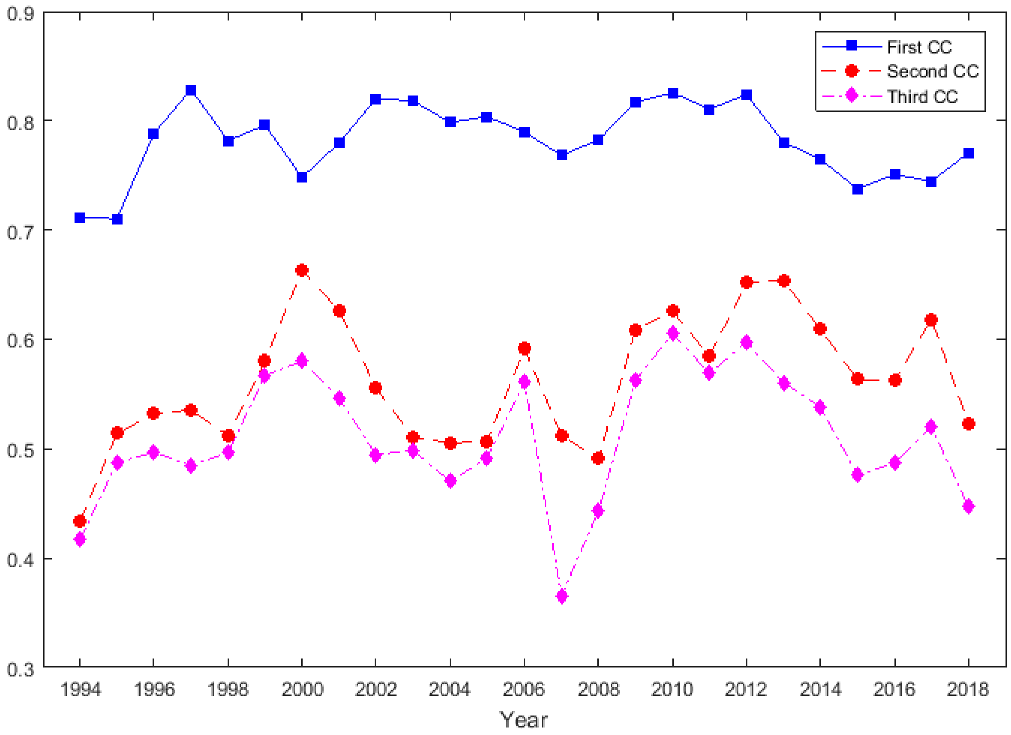

Figure 4. The first three canonical correlations between returns and order flows over time (Jan.1994 to Dec.2018)

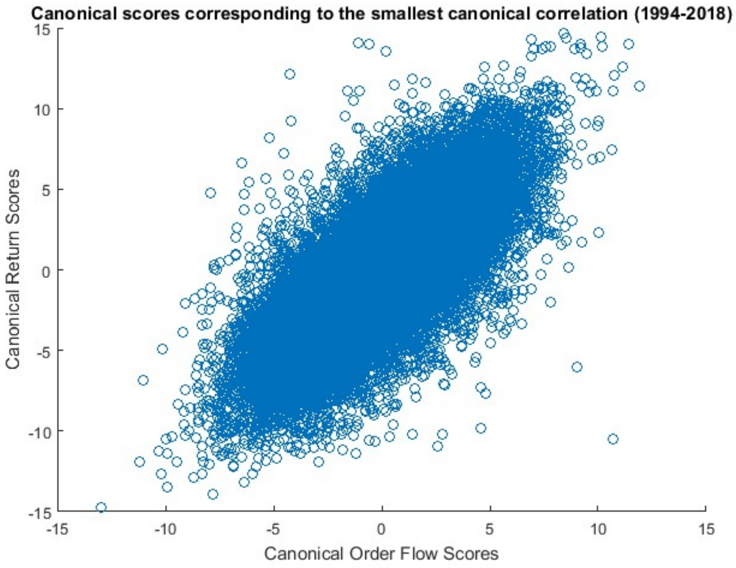

Figure 5. Canonical scores based on the smallest canonical correlation in model (11)

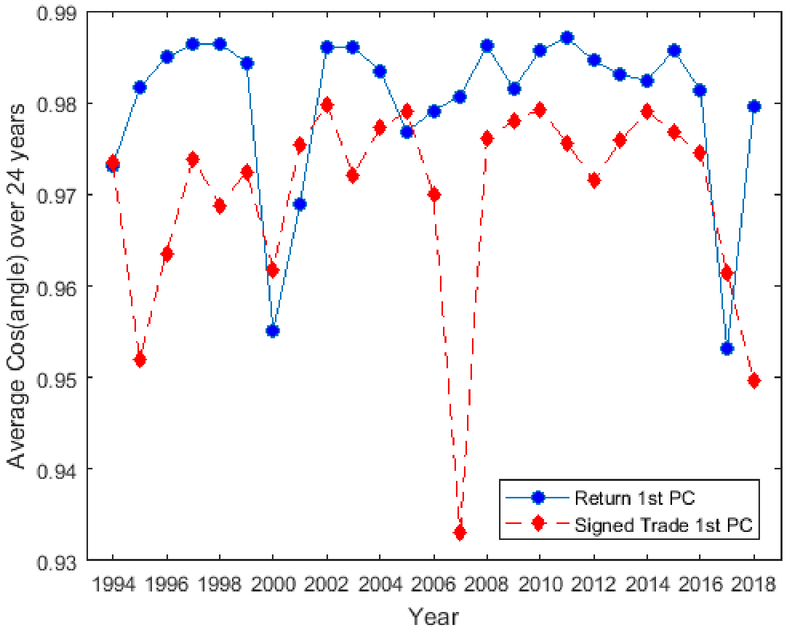

Figure 6. Similarity measure for the first PC of returns and order flows (Jan.1994 to Dec.2018)

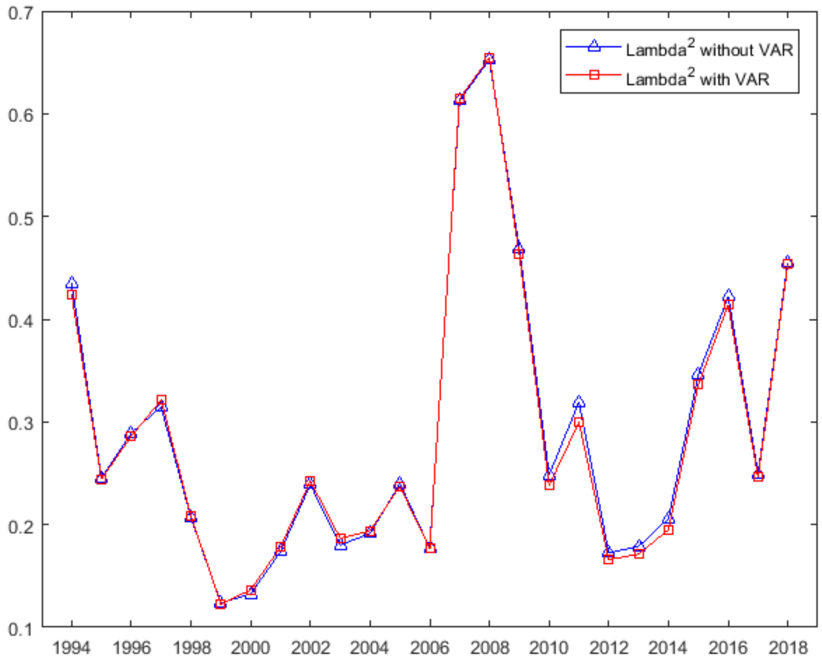

Figure 7. Index (10) based on models (7) and (12) (Jan.1994 to Dec.2018)

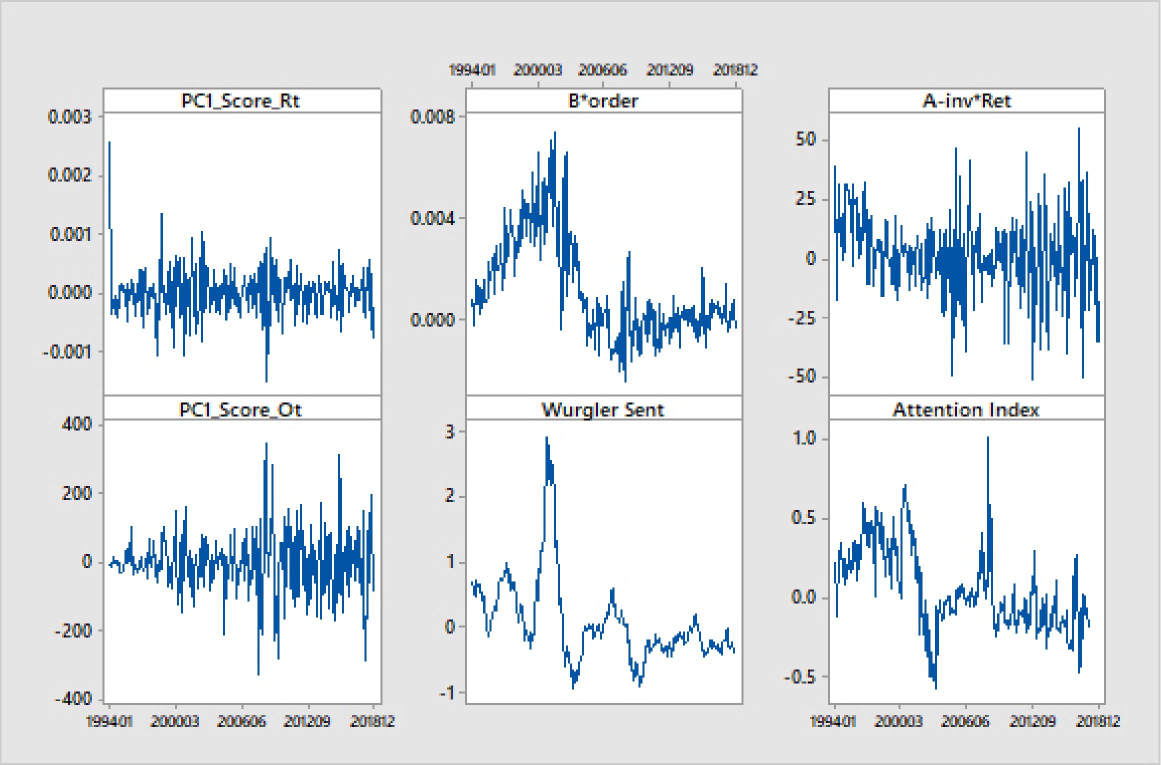

Figure 8. Time series plots of sentiment and commonality indices (from 1994 to 2018)

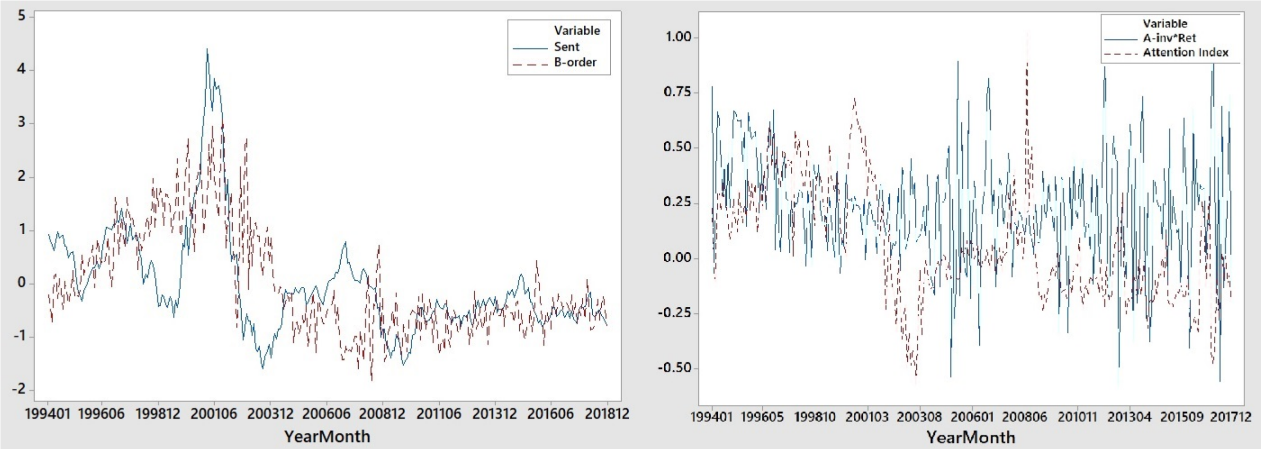

Figure 9. Time series plots of sentiment, attention and commonality indices (from 1994 to 2018)

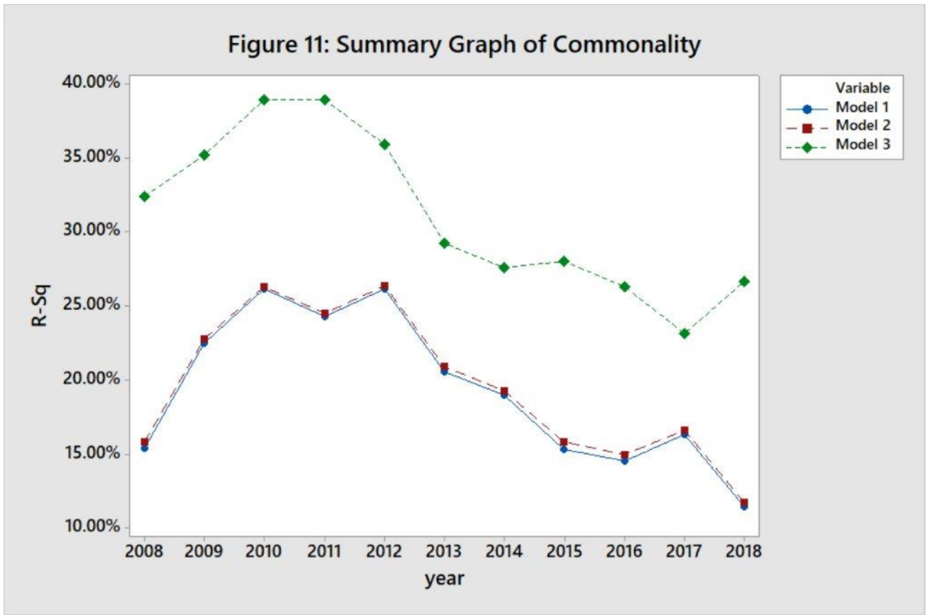

Figure 10. R^2 of firm level multivariate regressions on sentiment and attention (from 2010 to 2018)

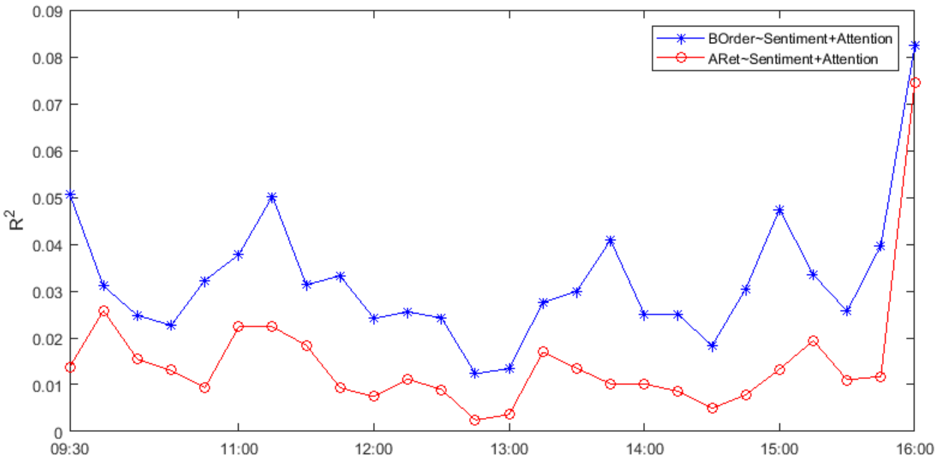

Figure 11. Intraday pattern of commonality, attention and sentiment relationship