The Information Content of Limit Order Books

Colin Swaney

September 25, 2016

ABSTRACT

Preliminary draft.

Recent changes in market microstrucutre have transformed the role of limit orders and the importance of limit order book data. This study contributes to a line of research assessing the information content of electronic limit order books by studying a unique sample of NASDAQ-traded limit order book data. I analyze the shape of limit order books and find that intraday variation in limit order book shapes is primarily explained by just two to three factors. Interestingly, these factors are also shown to predict intraday price movements. In further analysis I show that high-frequency trade is associated with significant decreases in the magnitude of responses to order book events. Overall, my results demonstrate that a small number of factors are sufficient to capture important features of order book dynamics.

1 Introduction

Markets, and the exchanges through which they operate, have experienced major changes in the last decade. Regulatory changes have facilitated the fragmentation of markets. New competition has encouraged innovation in market design and technology. The search for liquidity in diffuse markets has led to a new role for speed, and the decline of traditional market making has led to a new importance for detailed order data. These structural and technological changes have been exploited by a new class of high-frequency traders. O'Hara, 2015 argues, correctly, that the domination of markets by high-frequency traders has led to a new high-frequency market microstructure. And with a new market microstructure comes new sources of data, as well as new challenges for researchers attempting to make sense of that data.

An essential characteristic of this new market microstructure is the rise of the limit order. Traditionally, limit orders played no informational role in microstructure, in which informed traders relied completely on market orders. This assumption is untenable in a high-frequency world where limit orders are critical to optimal order execution[1]. Practitioners now rely on real-time data-feeds of order-level data, which are provided by exchanges and used as inputs to algorithmic trading strategies and execution algorithms. The main goal of this research is to bring some understanding to this new data.

Analyzing this type of data is challenging. First, their is a technical challenge presented by the size and structure of ultra high-frequency datasets. Theses datasets are often timestamped to the nanosecond, resulting in datasets requiring several terabytes of storage. Furthermore, such data is often provided in a semi-structured format, requiring considerable preprocessing. At the very least it is usually necessary to reconstruct limit order books from underlying message data. Second, there the challenge presented by the structure and level of detail of these datasets. There are many possible representations of limit order data, and it's not clear whether we should be analyzing the limit order book, message data, or both. The technical challenge, it turns out, is not so great. The analysis challenge, on the other hand, is formidable. In this paper I consider a simple representation of the structure of limit order books, and see what, if any, useful information this representation contains.

In this paper I start by collecting high-quality limit order book data. NASDAQ provides access to a real-time view of their market through TotalView-ITCH. This service provides subscribers with a live stream of out-going message data from which full limit order books can be reconstructed. Professional traders utilize this data to determine order

[1]Maker-taker pricing futher reduces the transaction costs of limit orders, while strategies used by high-frequency traders increase the incentive of informed traders to camouflage their orders.

routing decisions by comparing the liquidity available in competing exchanges, and deciding between market and limit order placements. In this paper I analyze a broad sample of historical ITCH data. I reconstruct the limit order books of more than 100 stocks over the one-year period beginning January 1, 2013[2]. In the process of reconstructing the limit order book I obtain a complete message history for each of the stocks in my sample. In my analysis I make use of both sources of data.

Next I analyze the shape of the limit order books. Order books contain information about future price movements, but exactly which features of order books best represent that information is unclear. The approach I take in this paper is to start by determining features of the order book that are important in explaining the *shape* of order books. Since theory does not offer a clear direction, I take a statistical, data-driven approach and analyze limit order books using principal component analysis (PCA). In order to study the commonality in order book shapes across stocks and time, I analyze the stock-day average of PCA applied to my dataset. The results show that limit order books can be described by a small number of factors, with the first two (three) principal components explaining more than 55% (65%) of variation on average across stock-days. Further examination of the PCA results reveals that the principal component weights take on familiar shapes: level, slope, and curvature (Litterman and Scheinkman, 1991; Afonso and Martins, 2012; Clarke, 2015).

In the next section I propose a set of order book shape factors based on the average principal component weights. These factors provide (approximately) optimal descriptions of order book shapes. Similar to the asset-pricing literature, I assume that these statistical features represent the underlying economic factors driving order book dynamics. In that case we should expect these factors may be correlated with future price movements as well. I test this hypothesis by estimating a vector autoregressive model combining returns with the shape factors. The results indicate that all three factors are statistically and economically significant predictors of future returns at lags of up to five minutes. Interestingly, the magnitudes of the average loadings on the factors have the same ordering as the importance of the factor shapes in explaining variation in the shape of the book: level, then slope followed by curvature.

In the final section I analyze the effects of high-frequency trading (HFT) on order book dynamics. HFT has become an important part of market microstructure, both in theory and in practice. In empirical studies HFT is generally found to improve overall market quality (Boehmer et al., 2014; Hendershott et al., 2011; Hagströmer and Nordén, 2013; Hasbrouck and Saar, 2013). On the other hand, HFT has come under scrutiny from regulators and investors who claim that HFT engages in unfair or illegal practices that harm institutional and retail investors. I use daily order message histories to calculate stock-day HFT intensity. The measure of HFT intensity that I use is aimed primarily at capturing the footprints left in order books by strategies associated with HFT market making. My analysis shows that this type of HFT is associated with decreases in the magnitude of the coefficients in the linear predictive model, and that this effect is robust to the inclusion of a variety of controls.

2 Data

Empirical market microstructure has traditionally depended on intraday trade and quote data. This choice made sense given the prevailing economic theory: these are the only variables that play meaningful roles in classical theoretical models of price formation (e.g. Kyle, 1985; Glosten and Milgrom, 1985). Much has changed since the early days of market microstructure. One of the most important changes has been the adoption of fully electronic limit order books. Modern electronic limit order books are essentially implemented as continuous double-auctions in which there is little or no role played by designated market makers, thereby opening the door for new agents, such as the high-frequency traders that have largely replaced them. Exchanges run by machines have resulted in drastically more sanitary and transparent trading and data collection processes. Indeed, improving the accuracy and transparency of displayed market liquidity were key arguments in favor of electronic limit order books. On the one hand this has led to a new source of rich and accurate data for researchers. But it has also led to an explosion in the size and velocity of data. Limit orders now play a central role in the observable trade process, a fact that is reflected in contemporary theoretical models (e.g. Kaniel and Liu, 2006; Goettler et al., 2009; Rosu, 2009, 2015). Empirical research now often relies on several levels of limit orders book data, leading to datasets of increasing size and complexity.

A wide variety of order book data have been used in previous studies, and from these several stylized facts have emerged. First, limit orders are numerous, but typically quite small. For example, order flow for large-cap U.S. stocks is in the hundreds of thousands of messages per day (more than four messages per second) on the NASDAQ exchange alone, while for the State Street S&P 500 ETF it is several million messages per day (more than forty messages per second). Yet the median order size for many stocks is one round lot (100 shares), and a significant

[2]Obtaining more complete samples typically requires purchasing data. My sample is constructed from data that is made available to researchers under a nondisclosure agreement.

number of trades now involve odd lots (O'Hara et al., 2014). Second, little of the message data communicates trades. The vast majority of message data relays the addition and cancellation of orders, with as little as 5% of messages corresponding to executed orders. This feature of order book data is reflected in the preponderance of so-called "fleeting" orders (Hasbrouck and Saar, 2009)[3]. Finally, "walking the book", in which market orders partially execute at prices beyond the best bid/ask, is exceedingly rare even amongst stocks that are thin at their inner levels (e.g. GOOG). For some equities less than 1% of filled orders lie outside the top level of the order book (Hautsch and Huang, 2012), despite the fact that the books for these stocks are often extremely thin at the best bid/ask.

In this study I obtain limit order book data from the NASDAQ Historical TotalView-ITCH dataset. ITCH has several advantages compared to the order book data found in many studies. First, it covers equity markets that are the most relevant to financial research (NASDAQ and NYSE). Second, it contains message data from which order books can be reconstructed, as opposed to order book snapshots. The distinction is important because raw messages contain information that is potentially useful, and that cannot be recovered from order book snapshots. In this paper I demonstrate an example of this by using message data to measure high-frequency trade activity in a way that is impossible using order book snapshots. Lastly, unlike many of the data sources previously studied, ITCH data is widely available (it is freely available to academic researchers by NASDAQ under a nondisclosure agreement).

The Historical TotalView-ITCH dataset provides a historical record of out-going messages from the NASDAQ exchange that is identical to the TotalView-ITCH data feed NASDAQ provides to professional traders (for a fee). The dataset is updated each day with message data from the most recent trading day, with data availability starting from January 2, 2001. Each daily file contains a detailed sequence of messages generated by the NASDAQ system from which the visible limit order book of NASDAQ-traded stocks can be reconstructed. There are four types of messages: order messages, trade messages, event messages, and order imbalance messages. Event messages identify important market events, for example the start and stop of market hours and trading halts. Order book reconstruction is primarily based on the order messages, which consist of add orders, partial and full cancellation orders, and execution messages[4]. All messages are timestamped to the nanosecond.

One limitation of ITCH is that non-displayed add orders (i.e. hidden liquidity) are not shown in the message data, only the execution of orders against non-displayed orders. This does not materially affect the reconstruction of the order book. It simply means that the order books that I reconstruct are the same as the order books that would be observed in real-time by market participants, and that I am therefore unable to analyze the relative information content of hidden liquidity in the market. Hidden liquidity is generally not found in order book data, and there is so far little research on this topic. A notable exception is Beltran-Lopez et al., 2009, which analyzes a dataset of German stocks provided by Xetra that offers a complete view of the available liquidity[5].

A further limitation of my sample is that the ITCH messages only allow me to reconstruct the visible order book internal to the NASDAQ exchange, not the consolidated order book. Therefore the order book data in this study represents only a partial view of the national market. It is, however, a particularly large view: NASDAQ's market volume share is around 35% for NASDAQ-listed stocks in my sample, and approximately 15% for NYSE-listed stocks. In addition, to the extent that updates to the alternative exchanges impact the mid-price of the NASDAQ exchange, this limitation biases against my findings. It is also worthwhile to note that the consolidated order book may not be the relevant view of liquidity from a trader's perspective. This is due to the fact that Reg. NMS, Rule 611 ("order-protection") only protects level one of the order book, meaning that market orders are not guaranteed to execute at the best available price if they exceed the available liquidity at the best bid/ask. In other words, theoretically it is the complete *disaggregated* order book data that is required to optimize trade execution.

I reconstruct the limit order book up to the best fifty quotes on each side of the book for a collection of 111 stocks for each trading day between January 1, 2013 and December 31, 2013. The stocks are chosen to match those included in the NASDAQ HFT dataset used in prior studies on high-frequency trade (e.g. Brogaard et al., 2014). Nine of the original 120 stocks in the HFT dataset (BARE, CHTT, KTII, BW, RVI, PPD, KNOL, ABD, and GENZ) were delisted prior to 2013. The remaining sample consist of 56 NYSE-listed stocks and 55 NASDAQ-listed stocks. As shown in Figure 1, the sample consists of a diverse set of stocks having a wide range of average daily returns, volumes, sizes, and prices. To my knowledge this is the largest, most recent sample of limit order book data studied that is freely available to academic researchers.

In order to reconstruct historical limit order books, ITCH message data must be translated from an efficient binary-format into meaningful message sequences. During the reconstruction process I retain the complete history

[3]Fleeting orders are improvements to the best bid/ask that are deleted almost immediately (e.g. within 100 milliseconds), which have been attributed to algorithmic traders "pinging" for hidden liquidity. Xu, 2014 proposes a theoretical model in which high-frequency traders rationally exhibit this type of trading pattern without a pinging motive.

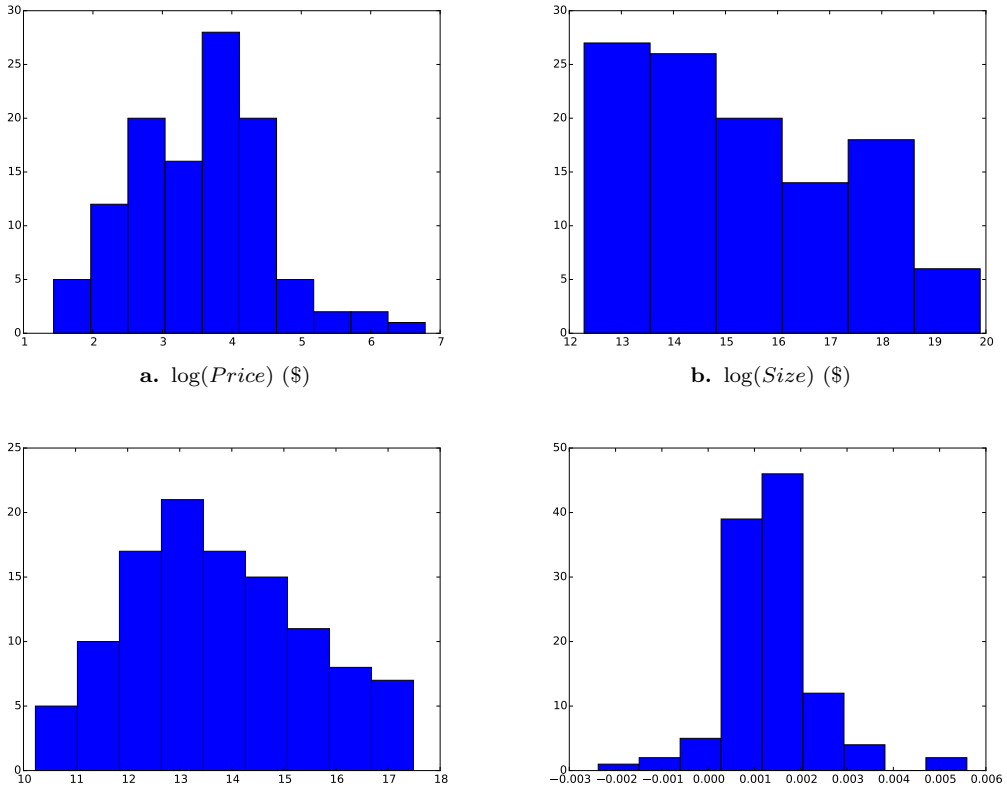
[4]There are also messages signaling that an order has been replaced by a new order with the same volume and side, but a different price. These messages are equivalent to a combination of a delete order and an add order with identical timestamps.

[5]Avellaneda et al., 2011 demonstrates the value of accounting for the *possibility* of hidden liquidity in predicting price movements, even though the true level is unobserved.

of relevant messages. The combined message and order book files are approximately 2.5 terabytes in total. In my main results I take snapshots of the order books at the end of each minute, resulting in 390 daily order book updates. I use the full message data in later analysis to construct daily measures of high-frequency trade intensity, which I combine with daily stock characteristics obtained from CRSP.

Figure 1: Distribution of stock characteristics.

This figure shows the distribution of daily average stock characteristics for my sample. $\log(Price)$ is the logarithm of the daily closing price; $\log(Volume)$ is the logarithm of the daily number of shares sold; $Return$ is the daily holding period return in basis points; $\log(Size)$ is the logarithm of the daily market capitalization in thousands. Averages for each stock are calculated from all days in the period Jan. 1, 2013 - Dec. 31, 2013 for which complete limit order book data is available.



a. $\log(Price)$ ($)

b. $\log(Size)$ ($)

c. $\log(Volume)$ (shares)

d. $Return$ (bp)

3 Order Book Shape

Prior literature has a proposed a variety of representations for limit order books, as well as features of the order book that might be informative about future price movements. From these studies two facts seem to be clear: that limit orders play an important role in price formation, and that this role is not restricted to the best bid and ask. The success of these varied features could, on the one hand, reflect the complexity of the price formation process. However, the poor motivation underlying many of these features suggests an alternative, and what seems more likely conclusion, that the order book dynamics are in fact driven by a small number of factors. Drawing on the asset-pricing literature, I take the view that the dynamics of the limit order book, and the price formation process in particular, are well-approximated by a low-dimensional linear model, and that the apparent success of the wide-range of proposed features is due to their correlation with these underlying factors. In this section I will demonstrate that such a parsimonious representation of the limit order book exists.

I represent limit order book data in a simple manner. First, I model the bid and ask sides of the book separately[6]. Second, with the exception of the prices of the best bid and ask (which are used to calculate mid-price returns), I ignore prices. Specifically, I don't use information about the difference in price between levels on a given side of the

[6]This decision is not entirely innocuous. It's possible that modeling the sides separately could ignore correlations between levels on opposite sides of the book, especially between the the best bid and ask.

Table 1: Variable descriptions.

This table describes the key variables used in this paper. Panel A contains descriptions of the variables used to model the limit order book. Panel B describes the variables used in the analysis of the effects of high-frequency trading on the limit order book dynamics.

Panel A: Order book variables	
Variable	Description
Ret	Mid-price returns computed from the average of the level-1 bid and ask prices at the end of each time period.
$Level^B(Level^A)$	The average of the volume of shares available for immediate execution at levels 1-10 of the order book.
$Slope^B(Slope^A)$	The average of the volume of shares available for immediate execution at levels 1-5 of the order book minus the average of levels 6-10.
$Curve^B(Curve^A)$	The average of the volume of shares available for immediate execution at levels 4-7 of the order book minus the averages of levels 1-3 and 8-10.
Panel B: Stock characteristics	
Variable	Description
HFT	The daily number of message runs as described in Hasbrouck and Saar, 2013 at the individual stock level. For a sequence of messages to be considered a run, the messages in a sequence must each be less than 100 milliseconds apart, there much be least 10 messages in the sequence, and the messages must be congruent (i.e. add messages must be followed by delete or execute messages, and delete messages must be followed by add messages)
$Price$	The stock-day closing price.
$Volume$	The daily number of shares sold at the individual stock level.
$Return$	The daily holding-period return at the individual stock level.
$Size$	The daily market capitalization at the individual stock level.

book, which reflects an assumption that the levels of the book are spaced (approximately) equally far apart. This allows me to model snapshots of the order book as N-vectors x_t, and the entire history of snapshots throughout a day as a $T \times N$ matrix X, where N is the number of levels of book data. This view of the order book can be thought of as an approximation to the "centered order book", in which the volume within each tick ($0.01) from the mid-price replaces the volume at each level. In practice these two representations are similar because I am only concerned with the dense part of the order book where gaps exceeding one tick are short-lived. Moreover, both representations will fail to capture any dependency of order book dynamics on the overall price level.

What do limit orders books actually look like? Figure 2 shows the average shape of a random selection of stocks in my sample. The average volume displayed at each level in Figure 2 is the simple average of the nanosecond time-stamped order book update in my sample for the given stock (i.e. the averages are not time-weighted). Order books display a wide-range of possible shapes. The overall average of my sample is quite similar to that of BRE. The liquidity in this order book is concentrated away from the best bid/ask, around the level-5 quote. The average displayed liquidity at all levels is quite small, between 2 and 3 round lots. The order book of PFE is typical of large-tick stocks: liquidity is concentrated towards the best bid/ask, and overall displayed liquidity is orders of magnitude larger than the other stocks shown. GOOG, on the other hand, is an extremely small-tick stock. Its order book is flat, with a pronounced increase at the best bid/ask. None of its levels exceed two round lots in average displayed liquidity. Interestingly, several of the stocks seem to have had strikingly asymmetric order books over the course of the year. For the most part, however, average order books shapes are symmetric.

Figure 2 also demonstrates that order books display a certain amount of similarity. I examine this similarity by applying principal component analysis (PCA) to limit order book data. PCA is a classical method of dimensionality reduction that has been used widely in finance and economics studies, and is therefore well-understood. Nonetheless, I briefly outline the details of this method in the context of the current application. The primary advantage of PCA is that it is a linear method. In particular, let X be a $T \times N$ matrix containing a sequence of order book snapshots over the course of a trading day. In general the dimension of X is $\min(T, N)$. In the present context the number of snapshots is much larger than the number of levels in the order book ($T >> N$), so I consider the rank of X to be N. The goal of PCA is to construct a low-dimensional approximation $X_k \approx X$, where $\text{rank}(X_k) = k < N$. The approximation can be expressed as $X_k = PW^T$, where P is $T \times N$, W^T is $N \times K$, and the rows of W^T are orthogonal.

The traditional way to perform PCA is to compute the eigen-decomposition of the sample covariance matrix $(X - \bar{X})^T(X - \bar{X}) = W\Lambda W^T$, and then define $P = (X - \bar{X})W$. In this case the eigenvalues (the diagonal of Λ) measure the contribution of the rows of W^T to the variance of X, and $X_k := P_k W_k^T$ is a rank k approximation of X explaining the greatest proportion of this variance. Alternatively, PCA can be viewed as solving the optimization problem

$$\min_{\text{rank}(X_k) \leq k} \|X - X_k\|_2,$$

the solution of which is given by $X = U_k \Sigma_k V_k^T$, where $U\Sigma V^T$ is the singular value decomposition of X, and U_k, Σ_k, V_k are matrices consisting of the first k columns of U, Σ, and V. It can easily be shown that $V^T = W^T$, and that $\Sigma^2 = \Lambda$. Either decomposition results in a representation of X in which each order book snapshot (i.e. a row x_t) is a linear combination of exactly k "typical" snapshots, which are the k eigenvectors $w_1, \ldots, w_k$ (i.e. the columns of W_k).

What do these typical shapes look like in my sample? In order to find out, I perform the following exercise. First, I compute averages of the complete, nanosecond time-stamped order book updates over one-minute intervals, resulting in one 390×10 order book matrix per side per stock-day. I then perform PCA on each of these stock-day matrices and compute stock-day averages of the eigenvector and eigenvalue matrices, $\bar{W}$ and $\bar{\Lambda}$, respectively. The results are shown in Figure 3. The top panel shows the average of the first three rows of W^T across stock-days in my sample. Not surprisingly, the bid and ask sides are symmetric. The loadings on the first factor are relatively evenly spread across the levels of the book, with the weight decreasing slightly towards the weights corresponding to the tenth level. The second factor places positive weight on the five highest levels of the order book (1-5), and negative weight on the five lowest (6-10). The third factor places negative weight on the highest (1-4) and lowest (8-10) levels, and negative weights on the levels in-between. The bottom panel shows the average eigenvalues associated with the principal components. This figure reveals that the first factor plays a dominant role in explaining the shape of the order book, explaining around 37% of the variation on either side of the book. The first three factors together explain approximately 65% the variation of order book shape.

Based on these results I propose three factors that I expect, due to their importance in determining the shape of the order book, also contain information about price movements. Borrowing from analogous analyses in the asset pricing literature, I refer to these factors as the "level", "slope", and "curvature" of the order book. I define the level as the simple average of the book. The slope I define as the average of the first five levels of the book minus the average of the last five levels of the book. The curvature is defined as the average of the middle four levels of

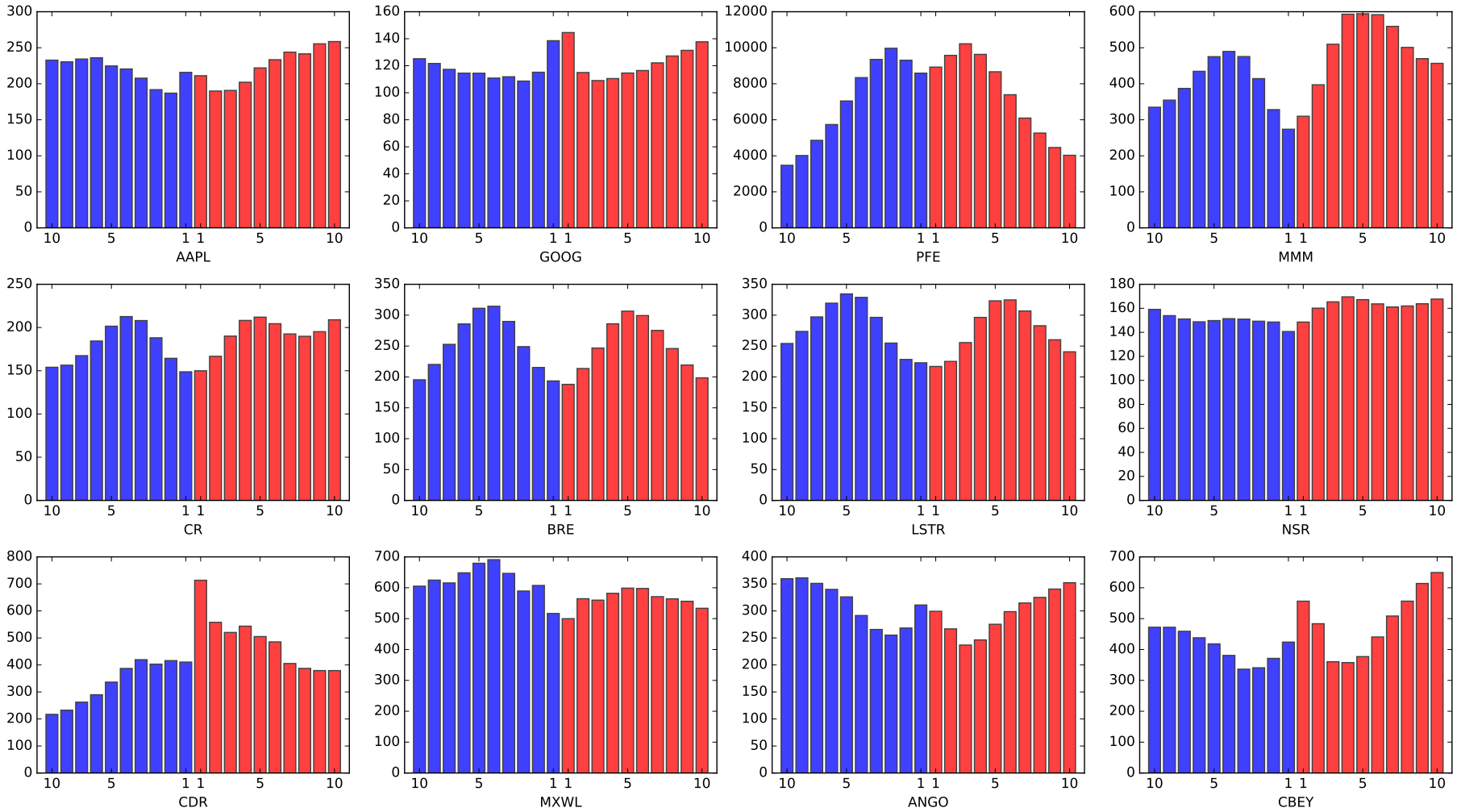
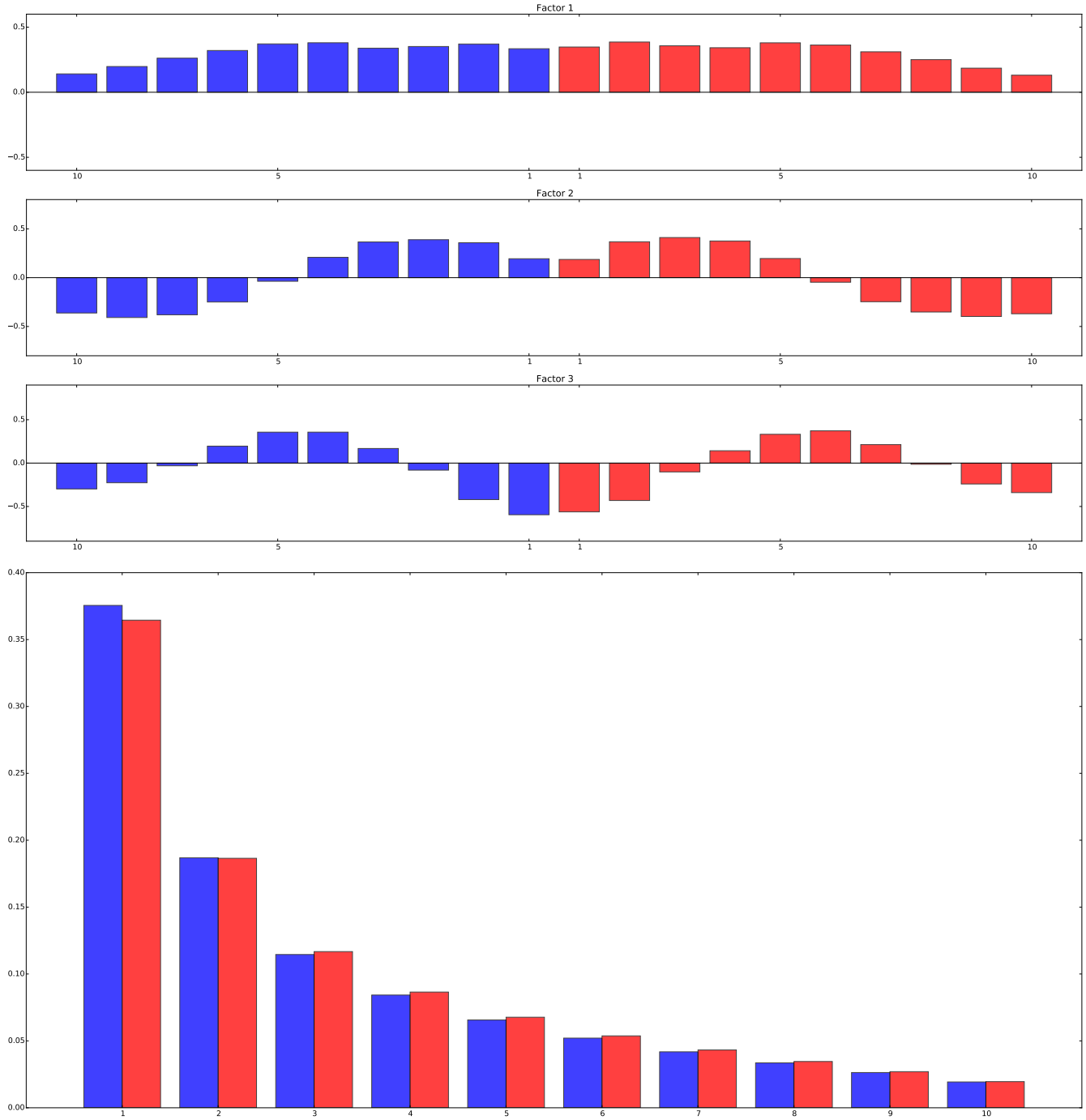


Figure 2: Examples of limit order book shapes.

This figure shows the average shape of randomly selected large (top row), medium (middle row), and small (bottom row) firms. Average shapes are calculated from all order book updates in the ITCH database during 2013.

Figure 3: PCA analysis of limit order books.

These figures show the results of the principal component analysis of daily limit order book volume data. For each day in the sample I perform PCA on the daily order book history. I then calculate the average of the stock-day eigenvectors (*top*) and the average of the stock-day eigenvalues (*bottom*). Averages are calculated from all order book updates in the ITCH database during 2013 for the stocks in my sample.



the book minus the average of the top three and bottom three levels of the book. The weights of these factors are designed to mimic the weights the first three eigenvectors shown in Figure 3. In the remainder of the paper I analyze the information content these proposed factors.

4 Price Prediction

In the previous section I identified a set of factors that explain a significant proportion of the intraday variation in order book shapes. In this section I examine whether these factors also contain information about the evolution of the order book. I am specifically interested in whether these factors are able to predict future price movements. At sufficiently short time horizons order books are mechanically related to returns in a linear fashion (Cont et al., 2013). At longer time horizons the order book may be connected to price movements either through liquidity demand or informed trade. As my proposed predictors are based on statistical factors, there is nothing to immediately identify them with either of these information channels, and I take no position on this issue. My concern in this study is whether the factors contain information, not what the source of that information is. The proposed factors might also be related to future changes in the shape of the order book itself. In that case the factor decomposition can be seen as a simplified alternative to modeling the evolution of liquidity at the tick level.

To answer these questions I choose another simple (linear) model, this time as an approximation of limit order book dynamics. Specifically, I estimate a vector autoregressive model of returns and first-differences in the proposed order book factors:

$$X_t = c + \sum_{k=1}^{K} A_k X_{t-k} + \epsilon_t, \tag{1}$$

where $X_t = \left(Ret_t,\ \Delta Level_t^{Bid},\ \Delta Level_t^{Ask},\ \Delta Slope_t^{Bid},\ \Delta Slope_t^{Ask},\ \Delta Curve_t^{Bid},\ \Delta Curve_t^{Ask}\right)$. VAR models are common in the market microstructure literature, where they have been used, for example, to compute information shares of competing channels (Hasbrouck, 1995). In this paper I view the VAR model as a linear approximation of order book dynamics, which are complex, nonlinear systems. (From this perspective the question this section addresses is whether the proposed factors capture meaningful features of that system). I use changes in factors instead of the factors themselves because the factors are highly persistent. Using first-differences reduces the number of lags required for consistent estimation of (1), and increases the efficiency of coefficient estimates by decreasing the correlation between regressors.

My primary interest is in the return equation

$$\begin{aligned} Ret_t = c + \sum_{k=1}^{K} & \beta_1 Ret_{t-k} + \beta_2 \Delta Level_{t-k}^{Bid} + \beta_3 \Delta Level_{t-k}^{Ask} \\ & + \beta_4 \Delta Slope_{t-k}^{Bid} + \beta_5 \Delta Slope_{t-k}^{Ask} + \beta_6 \Delta Curve_{t-k}^{Bid} + \beta_7 \Delta Curve_{t-k}^{Ask} \end{aligned} \tag{2}$$

Before discussing the results, it is useful to consider what signs are expected for the coefficient estimates in (2). The level factors unambiguously capture aggregate supply/demand, and I therefore expect to find a positive bid coefficient, and negative ask coefficient. The slope factor puts positive weight on levels that are closest to the best bid/ask, and negative weight on levels that are further away. Yuferova, 2015 argues that the inner portion of the book relates to same-side interest, while the outer portion of the book represents opposite-side interest of traders attempting to "lock-in" gains on directional bets. Since the inner and outer portions of the slope factor have opposite signs, the factor should be positively correlated with price movements for bids, and negatively with asks. An alternative view is that the slope simply captures shifts of supply/demand towards and away from the best bid/ask. Shifts towards the best bid (ask) signal future increases in price, while shifts away from the best bid signal that the current price is too high, thus this interpretation also leads us to expect a positive (negative) bid (ask) slope coefficient. For the case of the curvature factor I have no prior expectation.

I estimate equation (1) at the stock-day level with Ret_t given by the mid-price return between periods $t-1$ and t (i.e. the theoretical return of an asset with price equal to the average of the best bid and ask prices). I exclude stock-days for which I am unable to reconstruct a full time-series of returns and/or factors (e.g. due to trading halts), leaving a total of 25,715 stock-day observations. For each stock-day I take snapshots of the order book at one-minute intervals, giving 390 daily snapshots. Similar studies have tended to use intervals ranging between one and five minutes (Yuferova, 2015; Cao et al., 2009; Beltran-Lopez et al., 2009). I choose an interval of one minute and include five lags in equation (1), which is a typical number of lags selected by the AIC criteria in my sample. In Tables 2-5 I report the average of the estimated coefficients across stock-days, as well as t-values of the mean coefficients based on double-clustered standard errors.

Table 2: Linear model of the limit order book from shape factors.

This tables shows the average coefficient estimates of the vector autoregressive model of the limit order book: $X_t = \sum_{k=1}^{5} A_k X_{t-k}$. The components of X_t are one-minute mid-price returns and the changes to the level, slope, and curvature factors for each side of the order book. The table presents the stock-day average of coefficient estimate and t-values based on double-clustered standard errors. Coefficient estimates for the shape factors are scaled by 10^7. The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. ***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

	Panel A (dependent variable: Ret_t)				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Ret_{t-k}	-0.05***	-0.03***	-0.01***	-0.02***	-0.00
	(-5.80)	(-5.92)	(-4.82)	(-6.93)	(-0.91)
$\Delta Level^B_{t-k}$	3.87***	2.71***	1.96***	1.12***	1.01**
	(6.01)	(5.89)	(4.72)	(4.11)	(1.98)
$\Delta Level^A_{t-k}$	-4.09***	-2.62***	-1.90***	-1.36***	-0.76***
	(-6.62)	(-5.27)	(-5.22)	(-2.87)	(-2.67)
$\Delta Slope^B_{t-k}$	1.00***	0.76***	0.63***	0.51***	0.48***
	(5.18)	(4.57)	(4.05)	(4.12)	(3.47)
$\Delta Slope^A_{t-k}$	-1.20***	-0.81***	-0.76***	-0.61***	-0.36***
	(-5.27)	(-4.22)	(-4.70)	(-4.52)	(-3.54)
$\Delta Curve^B_{t-k}$	-0.61***	-0.24**	-0.20**	-0.11	0.03
	(-5.66)	(-2.02)	(-2.16)	(-1.51)	(0.26)
$\Delta Curve^A_{t-k}$	0.41***	0.41***	0.29***	0.24**	0.12**
	(5.26)	(4.14)	(4.52)	(2.86)	(2.06)

The results for the equation with return as the dependent variable are shown in Table 2. In agreement with prior studies, I find the sign of the lagged return coefficients are negative and statistically significant for all lags up to four minutes. The signs of the lagged level and slope factors on the bid (ask) side are all positive (negative), as expected, and the coefficients are statistically significant at all lags. The signs of the lagged curvature factors are the opposite of the level and slope factors for all but two of the coefficients (which are not statistically significant). For each of the factors I find that the estimated coefficients decrease in magnitude monotonically as the number of lags increases, and that the coefficients are approximately symmetric across the bid and ask sides of the book.

To give these coefficient estimates economic meaning I calculate the effect of a one standard deviation change in each of the independent variables on future returns. For the coefficient on one-minute lagged return this is equal to -4.5% of one standard deviation in returns. For the slope and curvature factors a one standard deviation change is associated with a future return of approximately 7%/9.5% and 6.5%/5% of one standard deviation in returns, respectively. For level factors the effects are somewhat larger: slightly over 12% for the bid side, and nearly 14% for ask side. Thus not only are the order book shape coefficients highly statistically significant, but they are also associated with economically significant effects on future returns. Interestingly, the ordering of these effects agrees with the ordering of the factors in terms of their importance in explaining the shape of the limit order book.

What about the dynamics of the order book factors themselves? Tables 3-5 show the results for the equations with level, slope, and curvature factors as the dependent variable, respectively. The results of the level equations show that, similar to returns, changes in the level factor are negatively autocorrelated, with the estimated coefficients decreasing monotonically as the number of lags increases. Not only are the level factors autocorrelated, but future changes on one side of the book are positively correlated with lagged changes in the level factor on the opposite side of the book. In fact, the magnitude of the autocorrelation is approximately equal to the magnitude of the cross-correlation. Similar results hold for the slope and curvature equations, except that the cross-correlation is much smaller than the autocorrelation in both equations. For all three of the factors I find statistically significant relations between lagged returns and future factor updates. For levels and curvatures lagged returns are negatively (positively) correlated with future changes on the bid (ask) side. For the slope equations the relationship is reversed.

At first glance the relationships seem mysterious. We can get some insight into these estimates by considering the implications of the factor construction process. The time-series for each of the proposed factors can be written as an inner-product of an order book history with a particular weight vector:

Table 3: Linear model of the limit order book from shape factors (continued).

This tables shows the average coefficient estimates of the vector autoregressive model of the limit order book: $X_t = \sum_{k=1}^{5} A_k X_{t-k}$. The components of X_t are one-minute mid-price returns and the changes to the level, slope, and curvature factors for each side of the order book. The table presents the stock-day average of coefficient estimate and t-values based on double-clustered standard errors. Coefficient estimates for the shape factors are scaled by 10^7. Coefficient estimates of returns are scaled by 10^4. The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. ***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

Panel B (dependent variable: $\Delta Level_t$)										
	Bid					*Ask*				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Ret_{t-k}	-0.85*	-1.71***	-1.43***	-1.13***	-1.08***	1.17***	0.97***	1.10***	1.08***	0.85***
	(-1.80)	(-3.22)	(-3.75)	(-3.94)	(-4.03)	(3.46)	(3.55)	(3.98)	(3.99)	(3.62)
$\Delta Level^B_{t-k}$	-0.186***	-0.11***	-0.09***	-0.06***	-0.02**	0.15***	0.12***	0.07***	0.06***	0.05***
	(-5.43)	(-4.67)	(-6.01)	(-5.38)	(-2.19)	(5.84)	(5.61)	(6.18)	(6.11)	(5.75)
$\Delta Level^A_{t-k}$	0.15***	0.12***	0.07***	0.05***	0.05***	-0.18***	-0.11***	-0.09***	-0.06***	-0.02**
	(5.92)	(5.83)	(6.51)	(6.41)	(5.85)	(-5.49)	(-4.69)	(-6.04)	(-5.42)	(-2.50)
$\Delta Slope^B_{t-k}$	-0.03***	-0.02***	-0.01***	-0.01***	-0.01***	-0.01***	-0.00	-0.00	-0.00**	-0.01***
	(-6.56)	(-4.76)	(-5.07)	(-4.95)	(-5.10)	(-3.90)	(-1.66)	(-0.50)	(-2.15)	(-3.29)
$\Delta Slope^A_{t-k}$	-0.01***	-0.00	-0.00	-0.00	-0.00***	-0.03***	-0.01***	-0.01***	-0.01***	-0.01***
	(-4.37)	(-1.26)	(-0.56)	(-0.88)	(-2.87)	(-6.38)	(-4.01)	(-5.01)	(-5.70)	(-5.48)
$\Delta Curve^B_{t-k}$	0.01***	0.00**	0.00	0.00	0.00	0.00	-0.00***	-0.01***	-0.00***	-0.00
	(5.07)	(2.25)	(0.27)	(0.036)	(0.62)	(0.82)	(-2.97)	(-4.10)	(-3.67)	(-1.73)
$\Delta Curve^A_{t-k}$	0.00	-0.00**	-0.00***	-0.00***	-0.000	0.012***	0.01***	0.00*	0.00	0.00**
	(1.52)	(-2.45)	(-3.97)	(-3.59)	(-0.78)	(5.76)	(3.66)	(1.69)	(1.36)	(2.07)

Table 4: Linear model of the limit order book from shape factors (continued).

This tables shows the average coefficient estimates of the vector autoregressive model of the limit order book: $X_t = \sum_{k=1}^{5} A_k X_{t-k}$. The components of X_t are one-minute mid-price returns and the changes to the level, slope, and curvature factors for each side of the order book. The table presents the stock-day average of coefficient estimate and t-values based on double-clustered standard errors. Coefficient estimates for the shape factors are scaled by 10^7. Coefficient estimates of returns are scaled by 10^4. The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. ***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

Panel C (dependent variable: $\Delta Slope_t$)										
	Bid					*Ask*				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Ret_{t-k}	2.37***	2.03***	2.34***	2.11***	1.95***	-2.06***	-3.06***	-2.99***	-2.76***	-2.13***
	(3.99)	(4.27)	(5.15)	(4.18)	(4.49)	(-4.19)	(-4.06)	(-5.04)	(-4.99)	(-4.33)
$\Delta Level^B_{t-k}$	0.11***	0.08***	0.02***	0.01	0.05***	0.14***	0.08***	0.04***	0.01**	0.04***
	(3.75)	(4.06)	(2.97)	(1.48)	(4.17)	(4.83)	(4.27)	(4.24)	(1.98)	(3.56)
$\Delta Level^A_{t-k}$	0.15***	0.10***	0.04***	0.02***	0.04***	0.11***	0.08***	0.03***	0.01**	0.04***
	(5.01)	(4.90)	(4.96)	(3.86)	(3.85)	(3.92)	(4.34)	(4.17)	(2.35)	(4.16)
$\Delta Slope^B_{t-k}$	-0.47***	-0.33***	-0.23***	-0.16***	-0.10***	-0.03***	-0.01***	-0.00**	-0.00	-0.01***
	(-8.36)	(-8.23)	(-8.18)	(-8.24)	(-8.35)	(-5.05	(-3.22)	(-2.29)	(-0.92)	(-4.39)
$\Delta Slope^A_{t-k}$	-0.03***	-0.01***	-0.00**	-0.00**	-0.01***	-0.46***	-0.32***	-0.23***	-0.16***	-0.09***
	(-5.48)	(-3.48)	(-2.41)	(-2.52)	(-4.60)	(-8.35)	(-8.22)	(-8.19)	(-8.27)	(-8.37)
$\Delta Curve^B_{t-k}$	0.02***	0.01***	0.00**	0.00	0.00	0.00	-0.00	-0.00**	-0.00***	-0.00
	(5.05)	(3.07)	(2.20)	(0.74)	(1.83)	(0.61)	(-0.94)	(-2.34)	(-3.16)	(-1.51)
$\Delta Curve^A_{t-k}$	0.00	-0.00	-0.00**	-0.00**	-0.00	0.02***	0.01***	0.01***	0.00*	0.00***
	(1.9)	(-0.38)	(-2.63)	(-2.63)	(-0.37)	(5.39)	(4.13)	(3.53)	(1.74)	(2.68)

Table 5: Linear model of the limit order book from shape factors (continued).

This tables shows the average coefficient estimates of the vector autoregressive model of the limit order book: $X_t = \sum_{k=1}^{5} A_k X_{t-k}$. The components of X_t are one-minute mid-price returns and the changes to the level, slope, and curvature factors for each side of the order book. The table presents the stock-day average of coefficient estimate and t-values based on double-clustered standard errors. Coefficient estimates for the shape factors are scaled by 10^7. Coefficient estimates of returns are scaled by 10^4. The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. ***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

Panel D (dependent variable: $\Delta Curve_t$)										
	Bid					*Ask*				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Ret_{t-k}	-0.55	1.88***	2.79***	1.61***	1.93***	-1.56*	-0.03	-0.97**	-1.21**	-1.11**
	(-0.33)	(2.90)	(3.29)	(3.27)	(3.01)	(-1.91)	(-0.03)	(-2.06)	(-2.47)	(-2.61)
$\Delta Level^B_{t-k}$	-0.26***	-0.19***	-0.11***	-0.07***	-0.10***	-0.21***	-0.15***	-0.08***	-0.04***	-0.08***
	(-5.20)	(-5.81)	(-6.00)	(-5.77)	(-5.47)	(-4.86)	(-5.61)	(-5.38)	(-4.54)	(-4.57)
$\Delta Level^A_{t-k}$	-0.23***	-0.16***	-0.07***	-0.05***	-0.08***	-0.23***	-0.17***	-0.09***	-0.06***	-0.09***
	(-5.03)	(-5.44)	(-5.44)	(-5.46)	(-4.58)	(-5.27)	(-5.82)	(-5.83)	(-5.57)	(-5.30)
$\Delta Slope^B_{t-k}$	0.05***	0.02***	0.02***	0.01***	0.02***	0.02***	0.00	-0.00	-0.00	0.00
	(6.09)	(4.92)	(4.72)	(4.02)	(4.96)	(3.52)	(0.93)	(-1.35)	(-1.62)	(0.74)
$\Delta Slope^A_{t-k}$	0.02***	-0.00	-0.01***	-0.01***	0.00	0.04***	0.02***	0.02***	0.01***	0.01***
	(3.7)	(-0.76)	(-2.86)	(-2.98)	(0.27)	(6.12)	(4.29)	(4.90)	(4.25)	(4.79)
$\Delta Curve^B_{t-k}$	-0.52***	-0.38***	-0.27***	-0.19***	-0.11***	0.00	0.01***	0.01***	0.00**	-0.00
	(-8.37)	(-8.27)	(-8.23)	(-8.30)	(-8.44)	(0.20)	(3.47)	(3.98)	(2.62)	(-1.18)
$\Delta Curve^A_{t-k}$	-0.00	0.00	0.01***	0.01***	-0.00	-0.52***	-0.38***	-0.27***	-0.19***	-0.11***
	(-0.80)	(1.76)	(2.85)	(3.40)	(-1.59)	(-8.39)	(-8.29)	(-8.26)	(-8.3)	(-8.46)

$$Level_t = Xf_1 := X\,(1,1,1,1,1,1,1,1,1,1)$$
$$Slope_t = Xf_2 := X\,(1,1,1,1,1,-1,-1,-1,-1,-1)$$
$$Curve_t = Xf_3 := X\,(1,1,1,-1,-1,-1,-1,1,1,1)$$

Similarly, the i^{th} principal component is equal to Xw_i, where w_i is the i^{th} eigenvector of the sample covariance matrix of X. The principal components are orthogonal by construction, a fact that follows from the orthogonality of the eigenvectors. In the previous section I computed the stock-day average of PCA eigenvectors. Averaging across samples does not preserve orthogonality. However, I find that the averaged eigenvector are in fact close to orthogonal. Moreover, the f_i I choose (modeled after the average eigenvectors) are also approximately orthogonal. As a result, the level, slope and curvature factors are designed to be approximately uncorrelated. On the other hand, limit order flow is known to be persistent (Biais et al., 1995). The combination of these facts suggests that the factors should predict themselves, but not the other factors. My results seem to reflect these observations.

5 HFT

I've shown that the shape of limit order books contain useful information concerning the price formation process. In this section I examine the role of high-frequency trading on the regression results from the previous section. Despite general acknowledgement that HFT plays a central role in modern markets (O'Hara, 2015), the effects of HFT on markets remains unclear. To begin with, there is the issue of defining HFT. HFT outfits are proprietary traders, and the opaqueness of their trading strategies leads to challenges not only in understanding how their activity impacts other traders, but in what the defining characteristics of HFT are. There are some obvious criteria, however. First, HFT trading strategies are implemented by machines. Second, HFT rely on technology that reduce the time required for information to pass between their machines and the machines of exchanges. Third, HFT trade frequently and make extensive use of limit orders (they are generally thought to have low trade to message ratios). Lastly, the trading strategies of HFT operate on intraday timescales, and HFT typically zero out their trading positions at the end of each trading day.

The adoption of electronic limit order books has predictably led to an increase in silicon trading, commonly referred to as algorithmic trading (AT). Authors have typically drawn a distinction between algorithmic trading and HFT, which is treated as a special case of AT. One view is that HFT primarily engage in market making strategies in which quotes are updated frequently in a balancing act between avoiding adverse selection and taking trade opportunities from competing HFT. Many electronic exchanges, in an effort to attract these traders, have adopted maker-taker policies that pay rebates to liquidity providers[7]. This view is consistent with reported volume shares of HFT trading, which fall in the range of 30-70%. In this paper it is this form of trading that I principally have in mind, although the measure of HFT intensity I use may capture other forms of AT.

Academic interest in HFT has primarily revolved around its impact on market quality, as measured by liquidity, efficiency and volatility. A common critique of HFT is that it does not provide liquidity during unfavorable market conditions, with the most striking example of this behavior being the evaporation of HFT liquidity during the 2010 flash crash (Kirilenko et al., 2015). Another critique of HFT is that it excels at anticipating the order flow of other traders, leading to high execution costs for non-HFT traders (Hirschey, 2013). Again, since HFT are generally not exchange-regulated market makers, they face no restrictions on liquidity consumption, and are therefore free to exploit these opportunities. On the other hand, as voluntary market makers, HFT possess no special access to markets (as traditional market makers do), except for that which they create through investment (e.g. co-location). What prevents non-HFT traders from exploiting this predictable order flow, given that HFT rely on the same data and technology that are available to other market participants? In fact, the majority of studies seem to support the conclusion that HFT has an overall beneficial effect on market quality[8].

In order to assess the impact of HFT on intraday return predictability, I regress the estimated coefficients from the VAR in the previous section on HFT intensity. Unlike the HFT dataset used in several prior studies, the ITCH dataset does not classify HFT. However, Hasbrouck and Saar, 2013 demonstrate that there is a simple method for constructing a measure of HFT intensity using unclassified message data that is highly correlated with HFT measures based on the labeled HFT dataset. Their method is based on counting the occurrence of simple patterns in message

[7]HFT market makers are generally not exchange-regulated, and therefore not obligated to provide liquidity. In fact, the spreads in many asset markets are such that pure market making strategies would not be viable without such rebates. For this reason this type of HFT is also referred to as "rebate traders."

[8]The two positions are not contradictory. For example, Lachapelle et al., 2015 shows in a mean-field game setting that the introduction of HFT agents may lead to improvements in the price formation process that are costly to non-HFT agents.

data that are characteristic of HFT market-making. Specifically, for each stock-day in my sample I count all of the sequences of messages having the following properties: (1) the messages are less than 100 milliseconds apart, (2) the messages correspond to orders on the same side of the book, (3) the orders referred to by the messages are all for same number of shares, (4) the messages are congruent in the sense that add orders are followed by delete orders (or an execute order, if it is the last message in the sequence), and vice versa, and (6) there are at least 10 messages in the sequence. While these properties may not be representative of all HFT trading strategies, they are typical of rebate-traders, and also unlikely to be produced by non-HFT. Since these sequence counts are correlated with overall message activity throughout the trading day, I normalize the raw sequence count by the daily volume of shares sold and use this value as the daily intensity of HFT.

Table 6 shows the results of regressing the VAR coefficients on HFT intensity and additional control variables. I consider the equation with return as the dependent variable, as our interest lies in HFT's impact on price formation. I run two versions of this regression. In the first version I use a dummy variable HFT_{top} equal to one for stock-days with HFT intensity above the median. In the second version, I take the logarithim of HFT intensity. In addition to the HFT intensity variables, I control for the daily closing price, the daily number of shares sold, daily holding period return, and market capitalization. Due to space limitations I only show the results for the first lag. The overall effect of HFT is to reduce the magnitude of the VAR coefficients. In the dummy variable version the effect is limited to the level coefficients, which are the most important of the VAR coefficients in the return equation. In the second version the HFT is also associated with dampened slope coefficients. Gould and Bonart, 2015 notes that the relationship between order imbalance at the top of the book and future returns depends on the tick size of the stock (i.e. the ratio of the minimum tick to the stock price). In contrast, I find that price has no effect on order book variables, although it does correlate with the lagged return coefficient.

6 Discussion

This paper contributes to a line of research studying limit order book data. Biais et al., 1995 is the earliest example of of this work. In that paper the authors document a number of important statistical properties describing limit order flow for the Paris Bourse. My paper is more closely related to Cao et al., 2009, Cenesizoglu et al., 2014, and Yuferova, 2015. Each of these papers investigates the information content of the limit order book beyond the best bid and offer, and find that variables related to higher levels of the order book predict intraday returns in-sample. In Cao et al., 2009 the authors analyze Australian Stock Exchange data and find that order imbalances measured at progressively higher levels of the book contribute modestly in predictive regressions of returns. In Cenesizoglu et al., 2014 the authors also estimate a vector autoregression model including variables describing the limit order book. The authors consider two types of variables. The first they call "depth", which is simply the number of shares available for immediate execution within some range of the order book levels. They also define "slope" variables, but they are unrelated to the slope variable defined in this paper. In contrast to my study, they construct their model in trade time, not physical time. Yuferova, 2015 is also quite similar to my study. In that paper the author investigates the information content of the inner and outer levels of order books using the Thomson Reuters Tick History database. My slope factor is similar to a linear combination of the inner and outer variables defined in that paper. This paper expands on Yuferova, 2015 by providing a statistical motivation for "slope-like" variables, as well as showing that these variables only provide a partial characterization of the order book and its information as it relates to future price movements.

In contrast with these studies, I show that the shape of limit order books can be described by a small number of statistical factors, each of which contain different information about future price movements. A similar approach is taken in Beltran-Lopez et al., 2009, but there are several key differences between our work. First, in Beltran-Lopez et al., 2009 the authors analyze price-impact functions constructed from limit order book data. In constrast, I show that commonality can be found in a direct representation of limit order books. Second, whereas the authors of the early work focus on commonality at the individual stock level, I identify commonality that exists across a broad sample of limit order books. Third, I show that common factors can serve as a low-dimensional model of the limit order book, and analyze how the parameters of that model depend on the presence of high-frequency trade.

As in most of the market microstructure literature, this paper does not consider out-of-sample performance. Gould and Bonart, 2015 demonstrates that a simple measure of limit order imbalance (using only level-1 information) predicts future price movements out of sample, especially for so-called "large tick" stocks. Given that in-sample R^2 are generally extremely low in intraday prediction models, it would be interesting to see what, if any, out-of-sample predictive power the proposed shape factors possess.

Table 6: The effect of high-frequency trade on predictive regressions.

This table shows the results of regressing the estimated vector autoregression coefficients on high-frequency trading: $Y_{i,t} = \alpha + \beta_1 HFT + \beta_2 \log(Price) + \beta_3 \log(Volume) + \beta_4 Ret_{i,t} + \beta_5 \log(Size) + \epsilon_{i,t}$. HFT is the the logarithm of the daily number of message runs (Hasbrouck and Saar, 2013) divided by daily volume. I include additional controls for daily price, volume, return, and market capitalization. HFT_{top} is a dummy variable equal to one if HFT is above the stock-day median. Due to space limitations I only show the effect on the coefficients with one lag from the vector autoregressive model. Standard errors are double-clustered. Coefficients for equations with level, slope and curvature as the dependent variable are multiplied by 10^7. The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. Data on common stocks are obtained from the CRSP database. HFT is computed from the ITCH message data. ***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively.

	Ret_{t-1}	$\Delta Level^B_{t-1}$	$\Delta Level^A_{t-1}$	$\Delta Slope^B_{t-1}$	$\Delta Slope^A_{t-1}$	$\Delta Curve^B_{t-1}$	$\Delta Curve^A_{t-1}$
$Intercept$	-0.22***	14.07**	-20.75***	6.86***	-9.82***	-4.40***	2.41***
	(-5.91)	(2.05)	(-5.54)	(3.57)	(-5.23)	(-3.38)	(2.86)
HFT_{top}	0.03***	-3.08***	2.07***	0.12	-0.30	0.00	0.03
	(3.606)	(-3.64)	(3.20)	(0.38)	(-1.21)	(0.02)	(0.20)
$log(Price)$	0.02**	0.22	-0.20	0.18	-0.11	0.28	-0.16
	(2.42)	(0.18)	(-0.35)	(0.71)	(-0.38)	(1.02)	(-1.22)
$log(Volume)$	0.03***	0.19	0.28	0.01	0.38	0.41	-0.14
	(5.02)	(0.13)	(0.52)	(0.06)	(1.36)	(1.26)	(-1.38)
Ret	-0.03	4.86	-6.06	4.96	2.80	2.65	-1.71
	(-0.35)	(0.34)	(-0.51)	(0.53)	(0.58)	(0.52)	(-0.57)
$log(Size)$	-0.02***	-0.78	0.82*	-0.44**	0.26	-0.18	0.03
	(-3.49)	(-0.69)	(1.80)	(-2.29)	(1.12)	(-0.67)	(0.26)
	Ret_{t-1}	$\Delta Level^B_{t-1}$	$\Delta Level^A_{t-1}$	$\Delta Slope^B_{t-1}$	$\Delta Slope^A_{t-1}$	$\Delta Curve^B_{t-1}$	$\Delta Curve^A_{t-1}$
$Intercept$	-0.36***	26.98***	-32.91***	3.59*	-7.14***	-3.70***	1.49**
	(-12.69)	(5.57)	(-10.20)	(1.80)	(-5.08)	(-4.26)	(2.18)
$log(HFT)$	0.01***	-1.21**	0.83***	0.37***	-0.28**	-0.14	0.09*
	(4.47)	(-2.57)	(3.84)	(2.93)	(-2.51)	(-1.42)	(1.66)
$log(Price)$	0.02**	0.20	0.49	0.37	-0.25	0.29	-0.10
	(2.59)	(0.16)	(0.95)	(1.34)	(-0.97)	(0.93)	(-0.78)
$log(Volume)$	0.04***	-0.57	1.62***	0.38	0.11	0.36	-0.02
	(7.24)	(-0.44)	(3.33)	(1.14)	(0.57)	(1.13)	(-0.24)
Ret	-0.02	5.84	-5.58	3.66	3.35	2.59	-2.05
	(-0.29)	(0.45)	(-0.47)	(0.38)	(0.66)	(0.54)	(-0.73)
$log(Size)$	-0.02***	-0.65	0.05	-0.71***	0.44**	-0.13	-0.06
	(-3.90)	(-0.58)	(0.11)	(-2.75)	(2.24)	(-0.47)	(-0.51)

7 Conclusion

In this paper I have attempted to provide a simple representation of electronic limit order books. My main insight is that if order book dynamics are primarily driven by a few factors, then we should be able to extract those factors from the limit order book shape. Using principal component analysis I show that three factors–limit, slope, and curvature– explain a majority of intraday variation in limit order book shape. Furthermore, these factors can be used to form a simple linear model of order book dynamics that predicts future price movements. High-frequency trading appears to dampen the relationship between prices and order book variables in this model, although the cause of this effect is unclear.

References

António Afonso and Manuel M F Martins. Level, slope, curvature of the sovereign yield curve, and fiscal behaviour. *Journal of Banking and Finance*, 36(6):1789–1807, 2012.

Marco Avellaneda, Josh Reed, and Sasha Stoikov. Forecasting prices from level-I quotes in the presence of hidden liquidity. *Algorithmic Finance*, 1:35–43, 2011.

Héléna Beltran-Lopez, Pierre Giot, and Joachim Grammig. Commonalities in the order book. *Financial Markets and Portfolio Management*, 23(3):209–242, 2009.

Bruno Biais, Pierre Hillion, and Chester Spatt. An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse. *Journal of Finance*, 50(5):1655–1689, 1995.

Ekkehart Boehmer, Kingsley Y. L. Fong, and Julie Wu. International Evidence on Algorithmic Trading. *SSRN working paper*, 2014.

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. High-frequency trading and price discovery. *Review of Financial Studies*, 27(8):2267–2306, 2014.

Charles Cao, Oliver Hansch, and Xiaoxin Wang. The information content of an open limit-order book. *Journal of Financial Markets*, 29(1):16–41, 2009.

Tolga Cenesizoglu, Georges Dionne, and Zhou Xiaozhou. Effects of the Limit Order Book on Price Dynamics. *SSRN working paper*, 2014.

Charles Clarke. The Level, Slope and Curve Factor Model for Stocks. *SSRN working paper*, 2015.

Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88, 2013.

Lawrence R. Glosten and Paul R. Milgrom. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100, 1985.

Ronald L Goettler, Christine A Parlour, and Uday Rajan. Informed traders and limit order markets. *Journal of Financial Economics*, 93(1):67–87, 2009.

Martin D. Gould and Julius Bonart. Queue Imbalance as a One-Tick-Ahead Price Predictor in a Limit Order Book. *SSRN working paper*, 2015.

Björn Hagströmer and Lars Nordén. The diversity of high-frequency traders. *Journal of Financial Markets*, 16(4): 741–770, 2013.

Joel Hasbrouck. One Security, Many Markets: Determining Contributions to Price Discovery. *Journal of Finance*, L(4):1175–1199, 1995.

Joel Hasbrouck and Gideon Saar. Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets*, 12:143–172, 2009.

Joel Hasbrouck and Gideon Saar. Low-latency trading. *Journal of Financial Markets*, 16(4):646–679, 2013.

Nikolaus Hautsch and Ruihong Huang. Limit Order Flow, Market Impact, and Optimal Order Sizes: Evidence from NASDAQ. In *Market Microstructure*, pages 137–161. John Wiley & Sons Ltd, 2012.

Terrence Hendershott, Charles M. Jones, and Albert J. Menkveld. Does Algorithmic Trading Increase Liquidity? *Journal of Finance*, 66(1):1–33, 2011.

Nicholas H. Hirschey. Do High Frequency Traders Anticipate Buying and Selling Pressure? *SSRN working paper*, 2013.

Ron Kaniel and Hong Liu. So what orders do informed traders use? *Journal of Business*, 2006.

Andrei Kirilenko, Albert S. Kyle, Mehrdad Samadi, and Tuzun Tugkan. The Flash Crash: The Impact of High Frequency Trading on an Electronic Market. *SSRN working paper*, 2015.

Albert S. Kyle. Continuous auctions and insider trading: Uniqueness and risk aversion. *Econometrica*, 53(6): 1315–1336, 1985.

Aimé Lachapelle, Jean-Michel Lasry, Charles-Albert Lehalle, and Pierre-Louis Lions. Efficiency of the price formation process in presence of high frequency participants: a mean field game analysis. *Mathematics and Financial Economics*, pages 1–40, 2015.

Robert Litterman and José Scheinkman. Common Factors Affecting Bond Returns. *Journal of Fixed Income*, 1(1): 54–61, 1991.

Maureen O'Hara. High frequency market microstructure. *Journal of Financial Economics*, 116(2):257–270, 2015.

Maureen O'Hara, Chen Yao, and Mao Ye. What's Not There: Odd Lots and Market Data. *Journal of Finance*, LXIX(5):2199–2236, 2014.

Ioanid Rosu. A dynamic model of the limit order book. *Review of Financial Studies*, 22(11):4601–4641, 2009.

Ioanid Rosu. Liquidity and Information in Order Driven Markets. *SSRN working paper*, 2015.

Jiangmin Xu. Optimal Strategies of High Frequency Traders. *SSRN working paper*, 2014.

Darya Yuferova. Intraday Return Predictability, Informed Limit Orders, and Algorithmic Trading. *SSRN working paper*, 2015.