EMPIRICAL FINANCE WITH LATENT STRUCTURE

by

Colin Swaney

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Business Administration
in the Graduate College of
The University of Iowa

May 2018

Thesis Supervisor: Associate Professor Artem A. Durnev

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Colin Swaney

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Business Administration at the May 2018 graduation.

Thesis committee: _____
Artem Durnev, Thesis Supervisor

_____
Jon Garfinkel

_____
Wei Li

_____
Qihang Lin

_____
Amrita Nain

This thesis is dedicated to my wife Xiayi and son Aedan.

ACKNOWLEDGEMENTS

I wish to thank Dr. Artem Durnev, Dr. Jon Garfinkel, and Dr. Qihang Lin for their help and encouragement. I also wish to thank Dr. Wei Li and Dr. Amrita Nain for their helpful suggestions and questions as members of my doctoral committee. Finally, I wish to thank my parents for their love, support, and patience.

ABSTRACT

This is the abstract.

PUBLIC ABSTRACT

This is the public abstract.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 INTRODUCTION 1

2 EVALUATING FUND MANAGER SKILL: A MIXTURE MODEL APPROACH 3

3 PRICE FORMATION AND ORDER BOOK SHAPES 4

3.1 Introduction 4
3.1.1 Summary 6
3.1.2 Contribution 8
3.2 Data 9
3.3 Order Book Shape 12
3.4 Order Book Dynamics 16
3.4.1 High-Frequency Trading 20
3.5 Conclusion 23

4 ORDER BOOK EVENTS ON A POISSON NETWORK 35

5 CONCLUSIONS 36

REFERENCES 37

APPENDIX 43

A LIMIT ORDER BOOK RECONSTRUCTION 43

LIST OF TABLES

Table

3.1 Variable descriptions 25

3.2 Vector autoregression estimates 26

3.3 The effect of high-frequency trade on predictive regressions 30

LIST OF FIGURES

Figure

3.1 Sample stock characteristics 32

3.2 Examples of average limit order book shapes 33

3.3 Principal component analysis of order books 34

CHAPTER 1
INTRODUCTION

My first essay considers that challenge of evaluating actively managed mutual fund managers. A key empirical issue is that of false discoveries: when evaluating thousands of funds, we are likely to find many funds that appear to deliever alpha even if the truth is that none actually do. Put simply, conventional p-values fail in this setting. I borrow from Barras et al. (2010) and suggest a method for classifying fund managers by imposing additional *structure* on the distribution of manager skill. By making assumptions about the distribution of manager skill, I am able to calculate probabilities concerning the skill level of managers. The results reveal that if the assumed structure is correct, then a large number of over-performing funds exists measured over long and short time horizons. However, consistent with earlier studies, trying to identify which funds will perform well in the future is not profitable.

In my second essay, I analyze Nasdaq limit order book. The issue I address is the information contained in the shape of limit order books. Empirical and theoretical studies agree that limit orders are are used by informed investors, and as a result, the shape of the order book should be related to future price movements. I show that order books have low-dimensional underlying structure, and that this structure predicts future price movements.

My third essay also analyzes Nasdaq limit order book data. In this case, I consider an event driven model of order book dynamics, as opposed to the state driven model in the preceeding chapter. The model directly estimates the connection

betweeen different types of order book events (limit orders, market orders, and cancellations) and distinguishes between endogenous events from exogenous events. The results highlight the importance of order book characteristics in explaining patterns in order arrivals.

An unintentional theme ties these essays together: each of the essays features a statistical model that depends on unobserved structure in some manner. In the first essay, I assume that the fund manager skill has a mixture of normals distribution; the second essay assumes that order books have hidden low dimensional linear structure; the final essay considers latent relationships between order book events. In each case, I use statistics to infer the latent structure, and then analyze the consequences for other aspects of the data.

CHAPTER 2
EVALUATING FUND MANAGER SKILL: A MIXTURE MODEL APPROACH

CHAPTER 3
PRICE FORMATION AND ORDER BOOK SHAPES

3.1 Introduction

Market orders are the traditional source of information in microstructure models. Market orders reveal informed traders' knowledge to market makers, who adjust quotes as they learn about underlying values. In modern exchanges, which rely heavily on informal, automated market makers, the roles of market and limit orders are less clear. Lacking the obligations and preferential treatment of specialists, voluntary market makers are free to demand liquidity as it suits them. A high-frequency market maker, for example, may periodically use market orders to opportunistically control her inventory (Xu (2015)). On the other hand, informed traders can minimize execution costs by supplying liquidity (Cont and Kukanov (2017)). As a result, the traditional views of the roles of market and limit orders are no longer adequate.

In fact, it is increasingly clear that limit orders are informative. This observation is supported by both theoretical (e.g. Kaniel and Liu (2006), Goettler et al. (2009), and Rosu (2009)) and empirical studies (e.g. Cao et al. (2009) and Brogaard et al. (2015)). In the present paper, I add to this evidence by examining the information content of limit orders submitted on the Nasdaq exchange. In particular, I show that the shape of the limit order book—i.e. the supply and demand curves generated by limit orders—is informative. My key findings are that a low-dimensional linear model explains most of the variation in the shape of order books, and that

the primary factors explaining the order book shape also predict returns. Thus, I characterize the way in which information presents itself through the shape of limit order books.

Limit order books are defined by pairs of vectors $p_t = (p_t^{(bid)}, p_t^{(ask)})$, and $v_t = (v_t^{(bid)}, v_t^{(ask)})$ containing the best N prices and shares available for trade at time t. The question I address in this paper is the relationship between v_t (the shape of the book) and the midpoint quote, $m_t = \frac{1}{2}\left(p_{t,1}^{(bid)} + p_{t,1}^{(ask)}\right)$. Specifically, I examine the relationship between the shape of order books and future midpoint returns. To do so, I follow a two-step process. First, I apply principal component analysis (PCA) to find a low-dimensional representation of the order book shape, $\tilde{v}_t$. Second, I model the time variation of the low-dimensional data, along with midpoint returns, using a vector autoregression (VAR) model. The overall result is a linear model of order book dynamics in which a handful of shape factors predict future price movements.

The following example motivates my approach. Suppose that the order book shape is generated by a small number of features $(w_1, \ldots, w_K)$, in the sense that each observation of the order book is a linear combination of the factors: $v_t = \sum_{k=1}^{K} f_k w_k$ for all $t = 1, \ldots, T$. In that case, the weight on each feature is given by its inner-product with the order book, $v_t' w_k$. Now suppose that the order book features predict future price movements such that $r_t = \sum_k \beta_k f_{k,t-1}$. Re-writing we have

$$r_t = \sum_k \beta_k \left(\sum_i v_{t,i} w_{k,i} \right) = \sum_i v_{t,i} \sum_k \beta_k w_{k,i} = \sum_i v_{t,i} \tilde{\beta}_i,$$

which demonstrates that the regression coefficients from a linear model using *all* levels of the order book as individual regressors are linear combinations of the regression

coefficients on the true features. The full regression generates the same predictions in this case, but it overlooks underlying structure in the data.

The approach can also be interpreted as a shrinkage (or de-noising) technique. PCA identifies a set of basis shapes that minimize the amount of "noise" in the data after projecting the original data onto them. Thus, the time series of factor coefficients can be regarded as de-noised versions of the original volume data, $\tilde{V}$. From this perspective, running a VAR model on $\tilde{V}$ instead of V amounts to training the model on "cleaner" inputs. Thus, the combination of PCA and VAR amounts to a two-step linearization of order book dynamics: the first step linearizes the input data, and the second step linearizes the dynamics.

3.1.1 Summary

I begin by constructing an extensive database of high quality limit order book data. Nasdaq provides access to a real-time view of its limit order books through the TotalView-ITCH database. This service provides subscribers with a live stream of out-going message data from which the full limit order book can be reconstructed. Using historical ITCH data, I reconstruct the order books of a broad sample covering more than one-hundred stocks over a one-year period. In the process of reconstructing order books, I obtain a complete message history for each of the stocks in my sample, which I use in further analysis to measure the prevalence of high-frequency trading.

Next, I analyze the shapes of the limit order books. I start by determining the features that best explain the shape of the order book. Since theory does not

offer a clear direction, I pursue a data-driven approach and obtain common shape factors by computing the average across stock-day results of PCA applied to order book snapshots. The results show that the first two (three) principal components explain more than 55% (65%) of the total variation in order book shape on average. Moreover, the principal components take on familiar shapes: they are the "level", "slope", and "curvature" of the order book.

In the following section, I propose a set of order book factors based on the level, slope, and curvature components. To the extent that the shape of the limit order book reflects underlying economic forces, I expect it to correlate with future price movements. It follows that the shape factors should contain information about future price movements. I test this hypothesis by estimating a vector autoregressive model combining returns with the shape factors. The results show 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 shape factors follow the same ordering as their importance in explaining the shape of the order book.

In the final section, I analyze the effects of high-frequency trading on order book dynamics. Empirical studies have generally found that high-frequency trading improves overall market quality (Boehmer et al. (2014); Hendershott et al. (2011); Hagströmer and Nordén (2013); Hasbrouck and Saar (2013)). I use order message histories to calculate a measure of daily high-frequency trade activity that aims to capture the footprints left in order books by strategies associated with high-frequency

market making. The results show that this type of high-frequency trading is primarily associated with decreases in the magnitude of the loadings on the level factor in the linear predictive model, suggesting that high-frequency market makers increase price efficiency.

3.1.2 Contribution

This paper contributes to a line of research studying the information content of limit order book data. It is closely related to Cao et al. (2009), Yuferova (2015), and Beltran-Lopez et al. (2009). Each of these three papers investigates the information content of the limit order book beyond the best bid and offer and finds that variables related to higher levels of the order book predict intraday returns. Cao et al. (2009) analyze data from the Australian Stock Exchange and find that order imbalances measured at progressively higher levels of the book contribute to predictive regressions of returns. Instead of focusing on individual levels, I recognize that there is commonality across levels of the order book and demonstrate predictability through a reduced set of common factors.

Yuferova (2015) investigates the information content of "inner" and "outer" levels of the order book using the Thomson-Reuters Tick History database. The slope factor I define is similar to a linear combination of the inner and outer variables defined in her paper. I extend Yuferova (2015) by providing a statistical motivation for including such variables, as well as showing that inner and outer variables only give a partial characterization of the order book and its information as it relates to

future price movements. More precisely, inner and outer variables omit the most important basis function of the limit order book: its *level*.

In contrast to these studies, I show that the shapes of limit order books are described by a small number of statistical factors, each of which contain distinct information about future price movements (the factors are orthogonal by construction). 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 order book data. In contrast, I show that commonality exists in a direct representation of the limit order book. Second, whereas the authors of that paper focus on commonality at the stock level, I identify commonality in the cross section of order books. Third, I show that these common factors predict future price movements in a low-dimensional model of order book dynamics, and examine how the parameters of that model depend on the presence of high-frequency trade activity.

3.2 Data

I obtain limit order book data from the NASDAQ Historical TotalView-ITCH database. This database provides a historical record of nanosecond time stamped messages transmitted by the Nasdaq exchange that is identical to data that professional traders purchase. The database divides into daily files containing sequences of binary messages that describe changes to the limit order book that market participants can feed into trading algorithms. Researchers can use these messages to

reconstruct snapshots of the limit order book throughout the day.

The database is massive: a single day of data contains hundreds of millions of messages, and a single year of data requires approximately one and a half terabytes of storage in its binary format (which is designed to reduce the amount of data passed between servers)—the database is several times larger after being converted into a practical format for research. Not only is the database large, but it is also disorganized. Each daily file is an exact copy of the messages transmitted to traders that day. Nasdaq passes all messages through a single channel, so messages for particular securities must be extracted from a sequence of messages for *all* securities traded on Nasdaq.

In addition, the message data must be translated from an efficient binary format into meaningful message sequences before reconstructing limit order books. During the reconstruction process, I record the complete history of relevant messages as well the state of the order book following each update. In my main results, I use snapshots of the top ten levels of the order book at the end of each minute to characterize order book shape, and I use the raw message data to calculate a measure of daily high-frequency trade intensity. I calculate daily stock characteristics using data from CRSP.

My main results rely on the reconstructed order books of a collection of 111 stocks for each trading day (9:30 am to 4:00 pm) between January 1, 2013, and December 31, 2013. The stocks are chosen to match those in a data set used in multiple recent studies of high-frequency trade (e.g. Brogaard et al. (2014)). Nine of

the 120 stocks in that data set (BARE, CHTT, KTII, BW, RVI, PPD, KNOL, ABD, and GENZ) delisted before 2013. The remaining sample consists of 56 NYSE-listed stocks and 55 Nasdaq-listed stocks. As shown in Figure 3.1, the sample contains a diverse set of stocks based on average daily returns, volume, size, and price.

ITCH data has several advantages compared to alternative sources of order book data. First, it covers equity markets that are the most relevant to financial research (Nasdaq and NYSE): other high-quality sources of order book data come from smaller, less active exchanges. Second, it contains message data describing every update to the limit order book, as opposed to order book snapshots. The distinction matters because messages provide information beyond the information in order book snapshots. For example, the calculation of the high-frequency trade activity metric in this paper requires more than order book snapshots. Lastly, Nasdaq makes ITCH data freely available to academic researchers under a nondisclosure agreement.

One limitation of ITCH data is that non-displayed add orders don't generate messages, and therefore we are only able to reconstruct the *visible* limit order book. This feature means that the order books I reconstruct are the same as the order books that market participants observe in real time. Hidden liquidity is usually not observed in order book data, and there is limited research on its importance. A notable exception is Beltran-Lopez et al. (2009), in which the authors obtain data from Xetra containing the entire order book. Avellaneda et al. (2011) suggests a method for approximating the level of hidden liquidity from trade events, but doesn't validate its procedure on real-world data.

A further limitation of ITCH data is that the messages only describe activity on the Nasdaq exchange, not the consolidated order book. Therefore, the order book data in this study only gives a partial view of the national market. It is, however, a substantial view: Nasdaq's market volume share is around 35% for Nasdaq-listed stocks in my sample and approximately 15% for NYSE-listed stocks. Also, to the extent that updates to the alternative exchanges impact the mid-price of the Nasdaq exchange, this limitation biases against the results of this paper.

3.3 Order Book Shape

Two important observations emerge from recent empirical work in market microstructure: first, that limit orders play a significant role in price formation, and second, that role is not restricted to the best bid and ask. Microstructure theory has arrived at the same conclusions, but connecting theory with exchange data is challenging: theory often makes simplifying assumptions that are difficult to reconcile with reality. As a result, we find a variety of proposed limit order book features that might predict future price movements, but little guidance in how to compare these features. Drawing on the asset-pricing literature, I take the view that a low dimensional linear model provides a reasonable approximation to actual limit order book dynamics and that the apparent success of a wide range of proposed features is due to their correlation with these underlying factors. In this section, I demonstrate that such a parsimonious representation of the limit order book exists.

Throughout this paper, I adopt a simple representation of limit order book

data. First, I model the bid and ask sides of the book separately. Second, with the exception of the prices of the best bid and ask (which are used to calculate mid-price returns), I ignore prices. This assumption allows me to model snapshots of the order book as N-dimensional vectors v_t, and the entire history of snapshots throughout a day as a $T \times N$ matrix V, where N is the number of levels of book data. This view of the order book is 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; the approximation tends to be better for stocks with small spreads.

Figure 3.2 shows the average shape of a random selection of stocks in my sample. The average volume displayed at each level in the figure is the simple average of the nanosecond time stamped order book updates (i.e. the averages don't take account of the duration between updates). We see that order books display a wide range of possible shapes. The overall average of the sample is similar to that of BRE; its liquidity is concentrated away from the best quotes, around the level-5 quote, and the average liquidity at all levels is small, between two and three round lots. The order book of PFE is typical of stocks with low price-to-tick ratios: liquidity concentrates towards the best quotes, and overall liquidity is orders of magnitude larger than the typical stock. GOOG, on the other hand, has a large price-to-tick ratio. Its order book is flat, and features a pronounced increase at the best quotes. None of its levels exceed two round lots in average displayed liquidity. Overall, average order book shapes are symmetric across bid and ask sides of the book.

I explore the determinants of the shape of limit order books by applying prin-

cipal component analysis to order book snapshots. Principal component analysis is a method of dimensionality reduction; instead of describing order books in terms of ten levels, we wish to describe it using two or three characteristic shapes. Principal component analysis is a linear method, so the resulting description will still take a simple mathematical form.

In the context of limit order book data, principal component analysis works as follows. Suppose that X is a $T \times N$ matrix of order book snapshots recorded over the course of a day. In general, the mathematical dimension of X is the minimum of T and N. In the present context the number of snapshots is larger than the number of levels in the order book ($T >> N$), so we consider the rank of X to be N. Principal component analysis attempts to construct a low-dimensional approximation $X_k \approx X$, where the X_k is a rank k matrix, and $k < N$; the approximation takes the form $X_k = PW^T$, where P is $T \times N$, W^T is $N \times K$, and the rows of W^T are orthogonal.

The traditional algorithm to perform principal component analysis is to compute the eigenvalue decomposition of the sample covariance matrix $(X - \overline{X})^T(X - \overline{X}) = W\Lambda W^T$, and then define $P = (X - \overline{X})W$. 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, principal component analysis 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 is represented by an optimal linear combination of k characteristic snapshots.

For the order book data in my sample, I find the k characteristic snapshots as follows. 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. Next, I perform principal component analysis on each of the stock-day matrices and compute stock-day averages of the eigenvector and eigenvalue matrices, $\overline{W}$ and $\overline{\Lambda}$, respectively.

Figure 3.3 shows the results. The top panel contains the average of the first three rows of W^T across stock-days in the sample. Not surprisingly, the bid and ask sides are symmetric. The loadings on the first factor are spread evenly across the levels of the book, with the weight decreasing slightly towards the tenth level. The second factor places positive weight on the five highest levels of the order book (levels 1 through 5), and negative weight on the five lowest (levels 6 through 10). The third factor places negative weight on the highest (levels 1 through 3) and lowest (levels 8 through 10) levels, and positive weight on the levels in-between. The bottom panel shows the average eigenvalues associated with the principal components. The eigenvalues reveal that the first factor plays a dominant role in determining the shape of the order book, explaining around 37% of the variation on either side of the book, while the first three factors combined explain approximately 65%.

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 similar analyses in the asset pricing literature, I refer to these factors as the "level", "slope", and "curvature" of the order book. The level factor is defined as the simple average across levels of the book; slope is defined as the average of the first five levels of the book minus the average of the last five levels of the book; curvature is defined as the average of the middle four levels of the book minus the average of the top and bottom three levels of the book. The weights of these factors mimic the weights of the first three eigenvectors shown in Figure 3.3. In the next section, I analyze the information content these proposed factors.

3.4 Order Book Dynamics

The factors identified so far explain a substantial proportion of the intraday variation in order book shapes. In this section, I examine whether these factors are also related to order book dynamics. Specifically, I assess the information content of these factors through their ability 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 impact price movements through exogenous liquidity demand or informed trading. As the proposed predictors are purely statistical factors, there is nothing to identify them with either of these channels, and I take no position on this issue. My concern in this study is whether the factors contain information, not what the underlying source of that

information is.

To answer this question, I model order book dynamics through a simple linear model. Specifically, I estimate a vector autoregressive model of returns and first-differences in the proposed order book factors:

$$X_t = c + \sum_{k=1}^{K} A_k X_{t-k} + \epsilon_t, \tag{3.1}$$

where

$$X_t = \left(Ret_t,\ \Delta Lvl_t^{Bid},\ \Delta Lvl_t^{Ask},\ \Delta Slp_t^{Bid},\ \Delta Slp_t^{Ask},\ \Delta Crv_t^{Bid},\ \Delta Crv_t^{Ask}\right).$$

I replace the factors by their first differences because the factors are highly persistent: using first differences reduces the number of lags required for consistent estimation of the model and increases the efficiency of coefficient estimates by decreasing the correlation between regressors.

Vector autoregression models are standard in the market microstructure literature, where they have been used, for example, to measure the price impact of trades (Hasbrouck (1991)). In this paper, I view the vector autoregression 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 important features of that system.

The most interesting equation in (3.1) is the return equation

$$\begin{aligned} Ret_t = c + \sum_{k=1}^{K} & \beta_{1,k} Ret_{t-k} + \beta_{2,k} \Delta Lvl_{t-k}^{Bid} + \beta_{3,k} \Delta Lvl_{t-k}^{Ask} \\ & + \beta_{4,k} \Delta Slp_{t-k}^{Bid} + \beta_{5,k} \Delta Slp_{t-k}^{Ask} + \beta_{6,k} \Delta Crv_{t-k}^{Bid} + \beta_{7,k} \Delta Crv_{t-k}^{Ask}, \end{aligned} \tag{3.2}$$

the coefficients of which reflect the predictability of returns based on changes to the shape of the order book. It is useful to consider what signs we might expected for the coefficient estimates in (3.2). The level factor unambiguously captures aggregate supply and demand, and we therefore expect to find a positive bid coefficient ($\beta_{2,k} < 0$) and a negative ask coefficient ($\beta_{3,k} > 0$). The slope factor places positive weight on levels that are closest to the best quotes, and negative weight on levels 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 ($\beta_{4,k} > 0$), and negatively with asks ($\beta_{5,k} < 0$). An alternative view is that the slope captures shifts of supply and demand towards and away from the best quotes. Shifts towards the best bid signal future increases in price, while shifts away from the best bid signal that the current price is too high, so this interpretation also leads to a positive bid coefficients.

Equation (3.1) is estimated at the stock-day level with Ret_t given by the midpoint return between periods $t-1$ and t. I exclude stock-days for that are missing observations 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 the end every one-minute interval, giving 390 daily snapshots per stock-day; related studies use intervals ranging between one and five minutes (Yuferova (2015); Cao et al. (2009)). The main results estimate a vector autoregression that includes five lags, a representative number of

lags selected by the AIC criteria across stock-days in the sample. In Table 3.2, 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.

Panel A of Table 3.2 shows the results for the equation with returns as the dependent variable. In agreement with prior studies, I find that the signs of the lagged return coefficients are negative and statistically significant for all lags up to four minutes. The loadings on lagged level and slope factors for bids (asks) 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, the estimated coefficients decrease in magnitude monotonically as the number of lags increases, and the coefficients are approximately symmetric across the bid and ask sides of the book.

To give these estimates economic meaning, I calculate the effect of a one standard deviation change in each of the independent variables on the return in the following period. For the coefficient on one minute lagged returns, this equals to -4.5% of one standard deviation in future returns. For the slope and curvature factors, a one standard deviation change on the bid (ask) side 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 approximately twice as large: slightly over 12% for the bids and nearly 14% for asks. Thus, not only are the order book shape coefficients statistically significant, but they are also associated

with economically significant effects on future returns. Interestingly, the ordering of the magnitudes of these effects agrees with the ordering of the factors importance in explaining the shape of the limit order book.

What about the dynamics of the order book factors themselves? Panels B-C of Table 3.2 show estimates of the equations with level, slope, and curvature factors as the dependent variable, respectively. The estimates for the level equation indicate 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 considerably smaller than the autocorrelation in both cases. For all three of the factors, I find statistically significant relations between lagged returns and changes to factors. For level and curvature, lagged returns are negatively (positively) correlated with future changes on the bid (ask) side. For the slope equations, the relationship reverses.

3.4.1 High-Frequency Trading

We have seen that the shape of limit order books predicts future price movements. In this section, I examine the role of high-frequency trading on this predictability. Despite a general acknowledgment that high-frequency trading plays a

critical role in modern markets (O'Hara (2015)), the effects of high-frequency trading on markets remain unclear.

Academic interest in high-frequency trading has primarily focused on its impact on market quality, as measured by liquidity, price efficiency, and volatility. A common critique of high-frequency trading is that it does not provide liquidity during unfavorable market conditions, with the most striking example of this behavior being the evaporation of high-frequency trading liquidity during the 2010 flash crash (Kirilenko et al. (2015)). Another critique of high-frequency trading is that it excels at anticipating the order flow of other traders, potentially leading to higher execution costs for slower traders (Hirschey (2013)). However, there is also ample evidence that high-frequency trading has an overall beneficial effect of market quality[1].

To assess the impact of high-frequency trading on intraday return predictability, I regress the estimated coefficients from the vector autoregression model in the previous section on high-frequency trading intensity. Unlike the Nasdaq high-frequency trading data used in prior studies, the ITCH dataset does not identify the orders of high-frequency traders. However, Hasbrouck and Saar (2013) demonstrates that a simple method for measuring high-frequency trading intensity using unclassified message data is highly correlated with high-frequency trading measures based on labeled data. Their method counts the occurrence of simple patterns in message data that

[1]In fact, the two positions are not necessarily contradictory. For example, Lachapelle et al. (2015) shows in a mean-field game setting that the introduction of high-frequency traders leads to an improved price formation process, but that the resulting process is costly to the slowest traders.

are characteristic of high-frequency market-makers. Specifically, for each stock-day I count the number of sequences of messages that satisfy the following properties: (1) the messages in the sequence 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 follow delete orders (or an execute order, if it is the last message in the sequence), and vice versa, and (6) there are at least ten messages in the sequence. While these properties may not be representative of all high-frequency trading strategies, they are typical of rebate traders and are unlikely to be produced by retail or institutional traders. Since this sequence count increases with overall message activity, I normalize the raw sequence count by daily volume of shares sold and use the resulting value as a measure of the daily intensity of high-frequency trading.

There is a distinction between general algorithmic trading and high-frequency trading. Algorithmic trading refers to any automated trading system. For example, an algorithmic trader might automatically buy or sell assets based on statistical arbitrage opportunities, or other technical signals. Trading signals may or may not occur at high-frequency, but the algorithms response will be fast. In contrast, high-frequency traders primarily engage in market making strategies involving frequent quote updates in a balancing act between avoiding adversely selected trades and capturing trading opportunities from competing high-frequency traders. Some electronic exchanges have adopted so-called "maker-taker" pricing schemes that pay rebates liquidity providers to attract these types of traders, as the spreads in some asset mar-

kets are now small enough that pure market making strategies would be unprofitable without these rebates. The measure of high-frequency trading I use is designed to measure this type of trading.

Table 3.3 shows the results of regressing the VAR coefficients on high-frequency trading intensity and additional control variables. I only analyze the equation with return as the dependent variable, as our interest lies in high-frequency trading's impact on price formation. There are two versions of the regression: the first version uses a dummy variable HFT_{top} equal to one for stock-days with high-frequency trading intensity above the median; the second version uses the logarithm of high-frequency trading intensity. In addition to the high-frequency trading intensity, 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 lagged coefficients. In the dummy variable version, high-frequency trading is associated smaller (in magnitude) level coefficients; in the second version, high-frequency trading is also associated with dampened slope coefficients. As the level coefficients are the most important in terms of explaining future returns, the overall effect is a reduction in the immediate price impact of increases to these factors.

3.5 Conclusion

In this paper I have attempted to provide a simple representation of electronic limit order book dynamics. My main insight is that if order book dynamics are primarily driven by a few factors, then we might be able to extract those factors from

the limit order book shape. Using principal component analysis I show that three factors—level, slope, and curvature— explain the 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.

Table 3.1: Variable descriptions

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	The average of the volume of shares available for immediate execution at levels 1-10 of the order book.
Slope	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	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.

Note: The 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.

Table 3.2: Vector autoregression estimates

	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)

Note: The 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.

Table 3.2 (continued)

	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)

Note: 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.

Table 3.2 (continued)

	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)

Note: The 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.

Table 3.2 (continued)

	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)

Note: The 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.

Table 3.3: The effect of high-frequency trade on predictive regressions

	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)

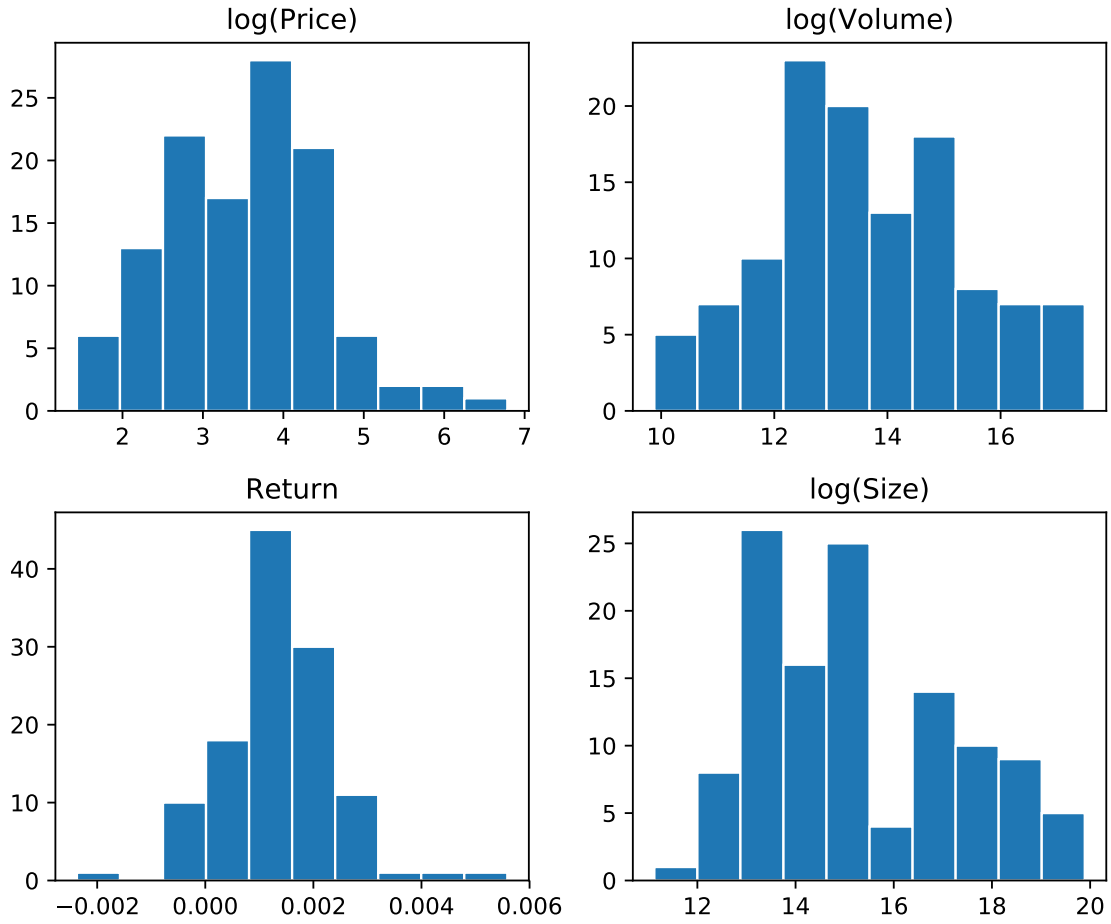
The table shows the results of regressing the estimated vector autoregression coefficients on high-frequency trading: $Y_{i,t} = \alpha + \beta_1 HFT + \beta_2 \log(Price) + \beta_3 \log(Volume) + \beta_4 Ret_{i,t} + \beta_5 \log(Size) + \epsilon_{i,t}$. HFT is the the logarithm of the daily number of 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. 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.

Table 3.3 (continued)

	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)

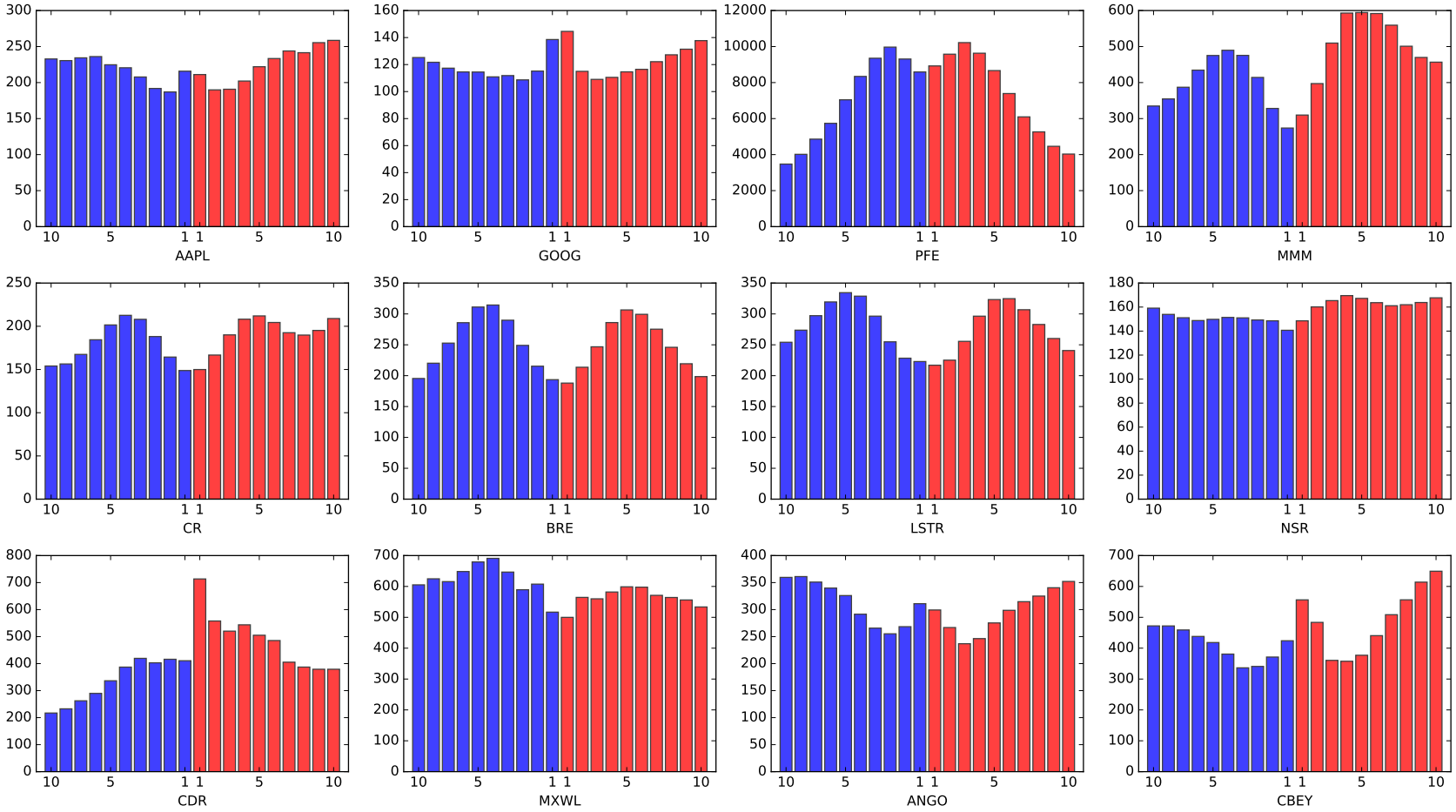
The table shows the results of regressing the estimated vector autoregression coefficients on high-frequency trading: $Y_{i,t} = \alpha + \beta_1 HFT + \beta_2 \log(Price) + \beta_3 \log(Volume) + \beta_4 Ret_{i,t} + \beta_5 \log(Size) + \epsilon_{i,t}$. HFT is the the logarithm of the daily number of 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. 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.

Figure 3.1: Sample stock characteristics



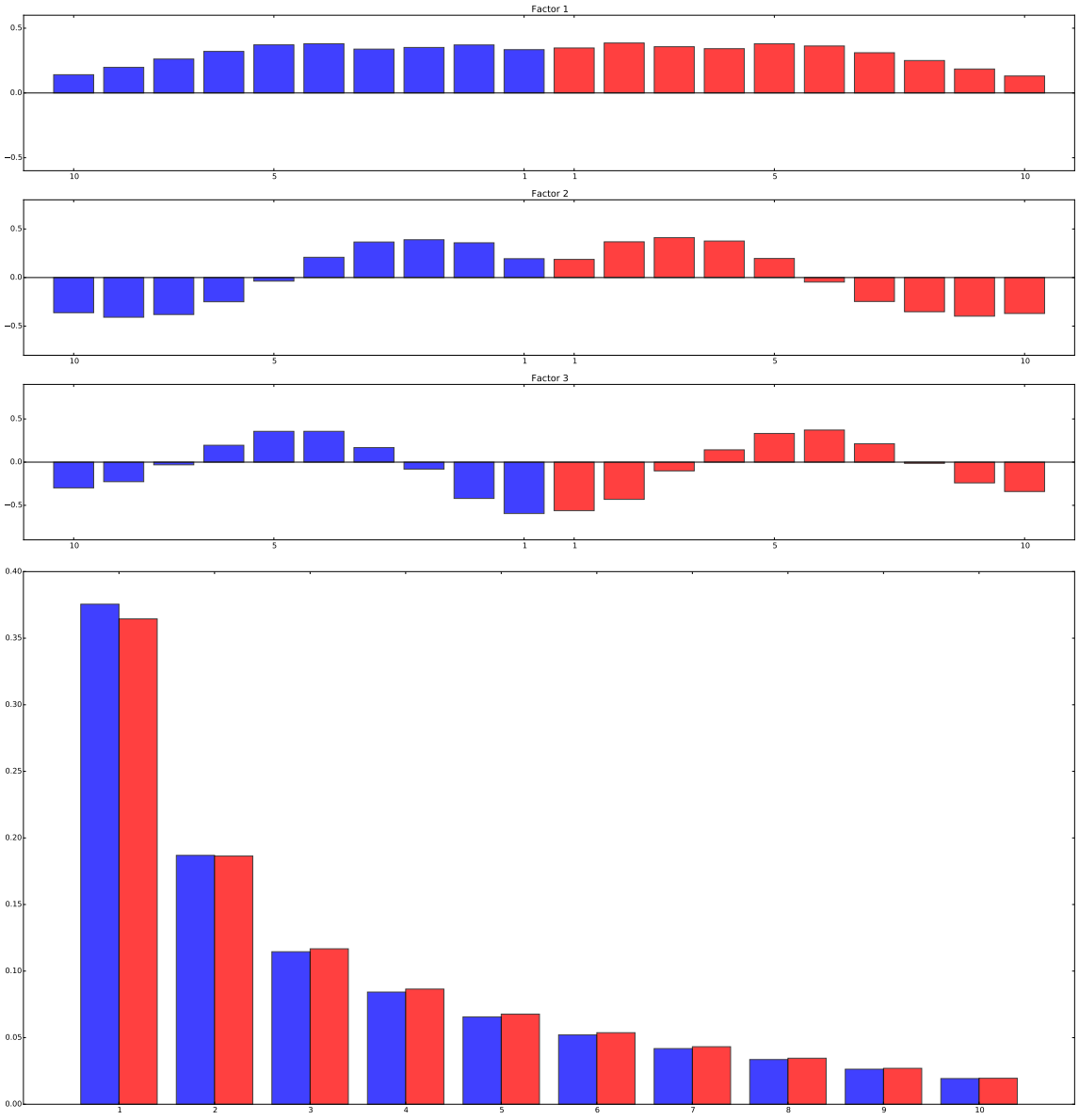
Note: These figures demonstrate the distribution of daily average stock characteristics for my sample. log($Price$) is the logarithm of the daily closing price; log($Volume$) is the logarithm of the daily number of shares sold; $Return$ is the daily holding period return in basis points; log($Size$) 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.

Figure 3.2: Examples of average limit order book shapes



Note: The 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 3.3: Principal component analysis of order books



Note: The 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.

CHAPTER 4
ORDER BOOK EVENTS ON A POISSON NETWORK

CHAPTER 5
CONCLUSIONS

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.

Aurélien Alfonsi and Pierre Blanc. Dynamic optimal execution in a mixed-market-impact hawkes price model. *Finance and Stochastics*, 20:183–218, 2016.

Robert Almgren and Niel Chriss. Optimal execution of portfolio transactions. *Journal of Risk*, 5(39), 2000.

Yakov Amihud and Haim Mendelson. Dealership market market-making with inventory. *Journal of Financial Economics*, 8:31–53, 1980.

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

Emmanuel Bacry, Sylvain Delattre, Marc Hoffman, and J.F. Muzy. Modeling microstructure noise with mutually exciting point processes. *Quantitative Finance*, 13(1):65–77, 2013.

Emmanuel Bacry, Thibault Jaisson, and Jean-François Muzy. Estimation of slowly decreasing hawkes kernels: application to high-frequency order book dynamics. *Quantitative Finance*, 16(8):1179–1201, 2016.

Walter Bagehot. The only game in town. *Financial Analysts Journal*, 27(2):12–14, 1971.

Klass P. Baks, Andrew Metrick, and Jessica Wachter. Should investors avoid all actively managed mutual funds? a study in bayesian performance evaluation. *The Journal of Finance*, 56(1):45–86, 2001.

Laurent Barras, Olivier Scaillet, and Russ Wermers. False discoveries in mutual fund performance: measuring luck in estimated alphas. *The Journal of Finance*, 65: 179–216, 2010.

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

Jonathan B. Berk and Richard C. Green. Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112:1269–1295, 2004.

Dimitris Bertsimas and Andrew W. Lo. Optimal control of execution costs. *Journal of Financial Markets*, 1:1–50, 1998.

Bruno Biais, Pierre Hillion, and Chester Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *The 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.

Jean-Philippe Bouchaud, J. Doyne Farmer, and Fabrizio Lillo. How markets slowly digest changes in supply and demand. In Thorsten Hens and Klaus Reiner Schenk-Hoppé, editors, *Handbook of Financial Markets: Dynamics and Evolution*, Handbooks in Finance, pages 57 – 160. North-Holland, San Diego, 2009. doi: https://doi.org/10.1016/B978-012374258-2.50006-3. URL http://www.sciencedirect.com/science/article/pii/B9780123742582500063.

Clive G. Bowsher. Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141:876–912, 2007.

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

Jonathan Brogaard, Terrence Hendershott, and Ryan Riordan. Price Discovery without Trading: Evidence from Limit Orders. *SSRN working paper*, 2015.

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.

Mark M. Carhart. On persistence in mutual fund performance. *The Journal of Finance*, 52:57–82, 1997.

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

Yong Chen, Michael T. Cliff, and Haibei Zhao. Hedge funds: the good, the bad, and the lucky. *Journal of Financial and Quantitative Analysis*, 52(3):1081–1109, 2017.

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

Rama Cont and Adrien de Larrard. Price dynamics in a markovian limit order market. *SIAM Journal of Financial Mathematics*, 4:1–25, 2013.

Rama Cont and Arseniy Kukanov. Optimal order placement in limit order markets. *Quantitative Finance*, 17(1):21–39, 2017. doi: 10.1080/14697688.2016.1190030. URL http://dx.doi.org/10.1080/14697688.2016.1190030.

Rama Cont, Arseniy Kukanov, and Sasha Stoikov. The price impact of order book events. *Journal of Financial Econometrics*, 0(0):1–42, 2013.

Maureen O'Hara David Easley, Marcos M. López de Prado. Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies*, 25(5):1457–1493, 2012.

A.P. Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39: 1–38, 1977.

Eugene F. Fama and Kenneth R. French. Common risk factors in the returns of stocks and bonds. *Journal of Financial Economics*, 33:3–56, 1993.

J. Doyne Farmer, Paolo Patelli, and Ilija I. Zovko. The predicitve power of zero intelligence in financial markets. In *Proceedings of the National Academy of Sciences of the United States of America*, February 2005.

Mark B. Garman. Market microstructure. *Journal of Financial Economics*, 3:257–275, 1976.

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

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

Martin J. Gruber. Another puzzle: The growth in actively managed mutual funds. *Journal of Finance*, 51:783–810, 1996.

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

Joel Hasbrouck. Measuring the information content of stock trades. *The Journal of Finance*, 46(1), 1991.

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

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.

Fumio Hayashi. *Econometrics*. Princeton University Press, Princeton, Princeton, 2011.

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

Patrick Hewlett. Clustering of order arrivals, price impact and trade path optimisation. In *Proceedings of the Workshop on Financial Modeling with Jump Processes*, 2006.

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

Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, 110(509):107–122, 2015.

Michael C Jensen. The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2):389–416, 1968.

Riadh Zaatour José Da Fonseca. Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. *The Journal of Futures Markets*, 34(6):548–579, 2014.

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.

Robert A. Korajczyk. High frequency market making to large institutional trades. Available at SSRN: https://ssrn.com/abstract=2567016, 2016.

Robert Kosowski, Allan Timmermann, Russ Wermers, and Halbert White. Can mutual fund "stars" really pick stocks? new evidence from a bootstrap analysis. *The Journal of Finance*, 61:2551–2595, 2006.

Albert S. Kyle. Continuous auctions and insider trading. *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.

Jeremy Large. Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10:1–25, 2007.

Scott W. Linderman. *Bayesian Methods for Discovering Structure in Neural Spike Trains*. PhD thesis, Harvard University, May 2016.

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

Hanno Lustig, Nikolai Roussanov, and Adrien Verdelhan. Common risk factors in currency markets. *Review of Financial Studies*, 24(11):3731–3777, 2011.

Albert J. Menkveld. High frequency trading and the new market makers. *Journal of Financial Markets*, 16:712–740, 2013.

Eli M. Remolona Michael J. Fleming. Price formation and liquidity in the u.s. treasury market: The response to public information. *The Journal of Finance*, 54(5):1901–1915, October 1999.

Iftekhar Naim and Daniel Gildea. Convergence of the EM algorithm for Gaussian mixtures with unbalanced mixing coefficients. In *International Conference on Machine Learning (ICML)*, 2012.

Yuriy Nevmyvaka, Yi Feng, and Michael Kearns. Reinforcement learning for optimal order execution. In *Proceedings of the twenty-third international conference on machine learning*, pages 673–680, 2006.

Whitney K. Newey and Kenneth D. West. Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61:631–653, 1994.

Jivendra Kale Nils H. Hakansson, Avraham Beja. On the feasibility of automated market making by a programmed specialist. *The Journal of Finance*, 40(1):1–20, March 1985.

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.

Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15:243–262, 2004.

Lubos Pastor and Robert F. Stambaugh. On the size of the active management industry. *Journal of Political Economy*, 120:740–781, 2012.

Marcello Rambaldi, Emmanuel Bacry, and Fabrizio Lillo. The role of volume in order book dynamics: a multivariate hawkes process analysis. *Quantitative Finance*, 17 (7):999–1020, 2017.

Christian Robert and George Casella. *Introducing Monte Carlo Methods with R*. Springer, New York, 2010.

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

Steven C. Mann Steven Manaster. Life in the pits: Competitive market making and inventory control. *The Review of Financial Studies*, 9(3):953–975, 1996.

John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society*, 64:479–498, 2002.

Hans R. Stoll Thomas Ho. Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9:47–73, 1981.

Gunther Wuyts. The impact of aggressive orders in an order-driven market: a simulation approach. *The European Journal of Finance*, 18(10):1015–1038, 2012.

Jiangmin Xu. Optimal strategies of high frequency traders. Available at SSRN: https://ssrn.com/abstract=2382378, 2015.

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

APPENDIX A
LIMIT ORDER BOOK RECONSTRUCTION