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

Table

LIST OF FIGURES

Figure

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

2.1 Introduction

The topic of this paper is evaluating the skill of actively managed equity mutual funds. There are two questions of primary interest. First, if we consider the population of mutual funds as a whole, is there any evidence that some funds are more skilled than others? In particular, are there any funds that we expect to produce returns that more than compensate their expenses? Second, supposing that such funds exist, can we identify these funds consistently enough to capture the value they provide? These questions are of obvious importance to investors choosing between actively and passively managed mutual fund investments.

In assessing fund manager skill, a critical issue is differentiating between skill and luck. Standard p-values are unsuitable criteria for measuring significance when performing multiple hypothesis tests because many of the tests are expected to reject the null hypothesis by chance. For example, in the case of actively managed equity mutual funds, the empirical distribution of performance is close to normally distributed, consistent with the hypothesis that all differences in performance arise entirely by luck. Recently, Barras et al. (2010) proposed a method to adjust for the existence of false positives based on a technique introduced by Storey (2002) in the context of biostatistical analyses. My paper an alternative perspective on the results found in Barras et al. (2010). My primary insight is that their framework suggests

an alternative estimation technique that enables a more comprehensive description of the distribution of fund performance. In contrast to Barras et al. (2010), this method leads to a direct and intuitive classification scheme, which I employ to assess the skill of individual funds. Additionally, my approach is more flexible in that it permits the possibility that the majority of managers generate returns that fail to justify their expenses. From a practical perspective, the alternative method I propose is preferable because it is much simpler to implement, and fast routines exist in several statistical programming languages.

My first set of results focuses on estimation. In this section, I employ the proposed methodology to describe the distribution of skill (*alpha*) amongst actively managed equity mutual funds. The results demonstrate that the collection of funds under consideration is consistent with a distribution in which more than 80% of funds generate slightly negative alpha, while minority populations of funds produce substantially positive or negative alpha. I label these minority populations the *Good* and *Poor* performers, respectively. These good and poor performing populations are approximately equivalent in size, but the mean performance of the poor performing population is twice that of the good performing population in absolute value. Unlike Barras et al. (2010), I find little difference between the distribution of long-run performance and the distribution of performance measured over a shorter time horizon.

In the machine learning jargon, my first set of results concern "unsupervised learning": there are no means of *directly* testing the quality of the statistical description. In my second set of results, I examine the performance of a simple investment

strategy to assess the quality of the parameter estimates indirectly. The investment strategy picks funds based on the unsupervised learning results, and its success depends on the stability of fund classifications over time. In fact, I find limited evidence to support persistence in fund performance: funds are unlikely to be classified as *Good* in consecutive years. When I analyze the performance of the investment strategy, I find that portfolios of *Good* funds generate positive alphas during my sample period, while portfolios of *Poor* managers generate negative alphas. However, in either case, the performance is similar to that of naive portfolios that invest in the top or bottom deciles of funds. Overall, the primary advantage of the classification system is that it frequently finds *no* positive alpha-generating population.

As a final experiment, I address the validity of the proposed methodology in the context of mutual fund analysis by performing a simulation exercise. The results show that the estimation procedure provides accurate estimates for plausible distributions of fund skill. The simulation also produces an intriguing finding: if we accept the model of Barras et al. (2010), and assume that their estimates are accurate, then the principal results of my investigation are improbable.

My work relates to Baks et al. (2001), which highlights the importance of investors' priors regarding the distribution of skill in determining allocations of wealth between actively and passively managed funds. Several papers address the issue of false discovery that complicates the task of formally evaluating the distribution of skill. For example, Kosowski et al. (2006) approaches the problem using a bootstrapping procedure. Their evidence suggests that the empirical distribution of fund

performance is unlikely to be observed through luck alone, and therefore skilled funds exist. More recently, Barras et al. (2010) attempted to account for false discoveries using a method introduced by Storey (2002) in the context of biostatistical analyses. Their study finds evidence of skill over short time horizons restricted to an early subsample of the data. They also determine that the size and performance of the skilled population have decreased over time.

2.2 Evaluating Skill

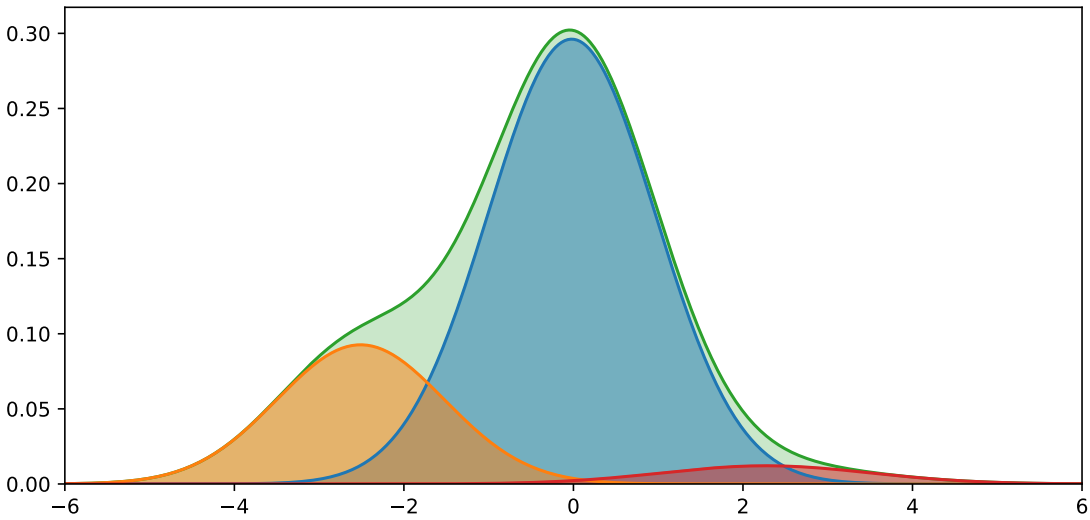
My strategy involves two parts. First, I propose a statistical model of the distribution of fund manager ability. The model assumes the existence of three distinct subpopulations of managers possessing varying degrees of skill: poor, average, and good managers. Second, I use an unsupervised learning method—an expectation maximization algorithm—to estimate the statistical model. Expectation-maximization gives a complete description of the distribution of manager skill and leads to an intuitive classification scheme for evaluating individual funds.

2.2.1 Framework for evaluating skill

The setting is that of an investor trying to evaluate mutual funds by their estimated alphas (relative to some asset pricing model). I assume that fund managers belong to one of three subpopulations of funds that possess different levels of skill: the differences in ability manifest themselves as differences in expected risk-adjusted returns across managers. Our investor observes the alpha generated by each fund, but not the subpopulation to which the fund belongs. Her first goal is to discover the

proportion of funds belonging to each subpopulation; her second goal is to predict the unobserved class of each fund. It is possible that she performs well on the first task, yet fails at the second, if the number of funds in a particular subpopulation is small, or if manager performance depends on unobserved, time-varying parameters. From an investment stand point, the first task may be important even if the later proves difficult because optimal asset allocation (from a Bayesian perspective) depends on an investor's prior beliefs about investment opportunities.

Figure 2.1: Mixture-of-normals distribution



Note: (Top) An example of a mixture model. The solid line represents the observed distribution; dotted lines represent the subpopulations. For the distribution shown, the proportions of the subpopulations shown come from Barras et al. (2010): 23% (red), 73% (black), and 4% (blue). (Bottom) The solid curve represents a kernel density estimate for the distribution of fund t-statistics found in the using the full sample of returns from 1975 to 2015; The dotted curve is the best fitting normal distribution fit to the same data.

The specific model I have in mind is the mixture of normals model depicted in Figure 2.1. With unconditional probability π_j, observation α_i is drawn from population j. Observations of alpha from population j have a normal distribution with mean μ_j and standard deviation σ_j. The skill level of each observation (denoted by z_i) is unobserved by our investor, and she is, therefore, unable to directly estimate the parameters μ_j, σ_j, and π_j $(j = 1, 2, 3)$. Referring to Figure 2.1, she observes the full distribution (the solid line), but she believes that the underlying populations exist and attempts to learn the parameters describing them.

2.2.2 Estimation procedure

I use an expectation-maximization algorithm to estimate the parameters of the model. Expectation-maximization is a numerical method used to obtain maximum-likelihood estimates of parameters in statistical models containing latent variables. For example, medical image reconstruction may involve the identification of unobserved tissue types from an f-MRI scan. The addition of a latent class variable makes the log-likelihood function impossible to maximize explicitly. Instead, an expectation-maximization algorithm provides an approximate numerical solution (that does not require the computation of derivatives) by replacing the maximization of likelihood by the maximization of *expected* likelihood, conditional on observed data and an evolving estimate of the underlying parameters.

Expectation-maximization algorithms arise in unsupervised learning tasks. The central idea is that many types of data contain unobserved classes that are

approximately normally distributed. Expectation-maximization provides a means of identifying classes by assuming their existence and identifying (in a statistical sense) a set of distributions that are most consistent with the observed data. Below I outline the basic algorithm; the appendix contains additional details.

Suppose that a statistical model of interest contains a set of observed variables $\mathbf{x}$, a set of latent variables $\mathbf{z}$, and a set of underlying parameters Θ, the value of which is unknown and must be estimated. Expectation-maximization works as follows. First, I choose an initial (coarse) estimate of the underlying parameters, denoted by $\Theta^{(0)}$. Second, the conditional expectation of the log-likelihood function $L(\Theta; \mathbf{x}, \mathbf{z})$ is computed, treating $\mathbf{z}$ as a random variable with a density conditional on both $\Theta^{(0)}$ (or $\Theta^{(k)}$ in step k) and a set of observations $\mathbf{x}$. Third, I choose $\Theta^{(1)}$ ($\Theta^{(k+1)}$) by maximizing the conditional expectation in the second step with respect to Θ. I repeat steps two and three until the estimate of Θ convergences (whenever $\Theta^{(k)}$ sufficiently close to the maximizer of the joint likelihood function). The resulting estimate is an approximate maximum likelihood estimate.

The mapping between this general description and the present application is as follows. First, the observed data $\mathbf{x}$ are mutual fund performances. Second, the latent variable $\mathbf{z}$ is a sequence of unknown fund abilities. Third, Θ consists of the means, standard deviations, and weights assigned to each skill population.

$$\Theta = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ \rho_1 & \rho_2 & \rho_3 \end{pmatrix}. \tag{2.1}$$

The expectation-maximization equations can be written down explicitly for the mixture of normals model. In more general settings, expectation-maximization requires

numerical evaluation of expectations (c.f. Robert and Casella (2010)). Moreover, the estimation procedure generates point estimates of the conditional probabilities that observation i comes from population j:

$$p_j^i = f_{y_i|x_i,\Theta}(y_i = j), \tag{2.2}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, J$. The estimates yield a simple classification scheme: each observation is identified with the maximum of the final estimated conditional probabilities evaluated at $\Theta = \hat{\Theta}$. Alternative classification methods can require that p_j^i exceed a specified threshold $\tau \in [0, 1]$.

Maximum likelihood estimators have useful properties under general conditions. First, maximum likelihood estimates are consistent, so we expect them to be close to the true parameter values given a sufficient sample size. Second, maximum likelihood estimates are asymptotically normal with theoretically known asymptotic covariance matrices. Third, maximum likelihood estimates are asymptotically efficient, meaning that Θ_{mle}^* minimizes the mean squared error amongst the set of consistent estimators (i.e. they achieve the Cramer-Rao bound).

Unfortunately, these results depend on asymptotic arguments: there is no general advice regarding the accuracy of these approximations for finite samples. Furthermore, the theorems that guarantee these properties rely on strong assumptions regarding the relevant data-generating process and, in the case of approximate maximum likelihood estimates, are sensitive to the choice of initialization. In the results that follow, I perform a simulation exercise that investigates the finite sample properties of the expectation-maximization estimates.

2.2.3 Why the EM algorithm?

The methodology in Barras et al. (2010) estimates the proportion of funds that are skilled, unskilled, or neutral, from which estimates of the false discovery rate are determined. If we accept that the mixture model suggested by the authors offers a good approximation to the true distribution, then it makes sense for us to try to learn the mean and standard deviation of the individual distributions as well. The expectation-maximization algorithms provide direct estimates of these statistics in addition to estimates of the proportions of the different populations. From these values, it is easy to determine the false discovery rate for a given significance level α. As I explain below, this perspective leads to a simple and intuitive classification scheme that can be used to construct portfolios of funds that are expected to consist of *Good*, *Bad*, or *Average* performers.

2.3 Data

I use the CRSP Survivor-Bias-Free Mutual Fund Database to identify open-end, diversified, actively managed mutual funds in existence between 1975 and 2014. I exclude funds designated as international, balanced, sector, bond, money market or index funds. From this collection of funds, I identify those with multiple share classes and aggregate their observations, with combined fund returns calculated as monthly TNA-weighted returns. I further restrict the sample by only including funds with at least sixty months of observations. I combine the resulting monthly returns data with Fama/French research returns data obtained via Wharton Research Data Services.

I use the returns data to construct a collection of regression intercept estimates (i.e. alphas) and corresponding t-statistics. To generate these values using the four-factor model studied by Carhart (1997):

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \epsilon_{i,t}, \tag{2.3}$$

where $r_{smb,t}$, $r_{hml,t}$, and $r_{mom,t}$ are the month t excess returns of factor-mimicking portfolios capturing size, value, and momentum premiums. I use t-statistics to measure skill because they account for differences in the precision of estimates and are asymptotically Gaussian. Computing t-statistics requires a choice of standard error computation: I compute standard errors according to Newey and West (1994), which adjusts for both heteroskedasticity and correlation across error terms.

Table 2.1 describes the returns data as well as the distribution of t-statistics. The data contains t-statistic observations from 2,831 funds. In agreement with prior studies, the mean alpha of an equally weighted portfolio of the funds over the full sample period is negative, and the model R^2 is close to 1 (e.g. Barras et al. (2010), Carhart (1997)).

2.4 Results

I begin by applying the expectation-maximization algorithm to the full data set of mutual fund returns. Next, I investigate the possibility of skill existing only over short time periods. I then analyze the distribution of fund performance over time and check for the existence of persistent skill. Finally, I validate my results through a simulation exercise.

Table 2.1: Summary fund return statistics

Period	$\hat{\alpha}$	$\hat{b}_{mkt}$	$\hat{b}_{smb}$	$\hat{b}_{hml}$	$\hat{b}_{umd}$	R^2
1975-2015	-0.5222	0.9906	0.2327	-0.0025	0.0204	0.9792
	(0.4814)	(0.0105)	(0.0312)	(0.0285)	(0.0174)	
1975-1980	-0.5562	1.0251	0.2623	-0.0705	0.1699	0.9833
	(1.1701)	(0.0223)	(0.0470)	(0.0206)	(0.0351)	
1980-1985	0.5616	0.9023	0.3392	-0.1827	0.0302	0.9845
	(0.7112)	(0.0177)	(0.0304)	(0.0290)	(0.0288)	
1985-1990	1.1940	0.9072	0.2749	-0.2193	0.0304	0.9953
	(0.5749)	(0.0169)	(0.0197)	(0.0380)	(0.0209)	
1990-1995	0.1405	0.9704	0.2660	-0.0920	0.0415	0.9929
	(0.4917)	(0.0123)	(0.0141)	(0.0169)	(0.0106)	
1995-2000	-1.4164	0.9649	0.2824	-0.0032	0.0006	0.9891
	(0.8081)	(0.0142)	(0.0205)	(0.0281)	(0.0195)	
2000-2005	-0.8233	1.0174	0.1925	0.1278	0.0144	0.9786
	(1.3075)	(0.0337)	(0.0232)	(0.0235)	(0.0242)	
2005-2010	-0.2280	1.0480	0.2161	-0.0864	0.0080	0.9894
	(0.7886)	(0.0352)	(0.0350)	(0.0414)	(0.0172)	
2010-2015	-1.8202	0.9837	0.2389	-0.0233	-0.0269	0.9928
	(0.6278)	(0.0173)	(0.0136)	(0.0270)	(0.0132)	
Min.	$q_{0.25}$	$q_{0.50}$	$q_{0.75}$	Max.	$\bar{\alpha}$	σ
-4.2230	-1.3690	-0.5542	0.2865	3.1210	-0.5509	1.2274

Note: The table reports the mean estimated coefficients and standard deviations of the Carhart (1997) four factor model for the years between 1975 and 2015 and for each five-year subperiod within that range. The bottom panel reports sample statistics and quantiels for the distribution of fund t-statistics using the full sample period.

2.4.1 Long-run skill

I start by estimating the parameters using t-statistics calculated from the complete time series of returns for each fund in the sample, as described in Section 2.3. Table 2.2 shows the results. The estimates contain several interesting findings. First, the proportions of good and poor performers are roughly equal, with each subpopulation making up approximately 8% of the overall population. The estimated

size of the best performing class is much larger than the corresponding estimate found in Barras et al. (2010) (8% v. 2.5%), while the predicted size of the poor performing class is significantly smaller (8% v 23%). Second, the estimates show that the population of average performing funds has a negative expected t-statistic. This observation calls in to question the assumption that the "typical" funds' performance is benign—investing in these funds is, in fact, detrimental in the long-run. Lastly, there is a substantial spread between the mean of the worst performing class and the best performing class. While the estimated difference in means refers to t-statistics, not returns, but it is still clear that there is a substantial cost to investing in poor performing funds.

2.4.2 Short-run skill

Next, I consider the evidence in support of manager skill over short horizons. Specifically, for each fund, I divide the time series of returns into five year sub-samples starting in the first year of the overall sample period, 1975. I consider each sub-sample as an individual fund and estimate its t-statistic over each sub-period for which the fund has at least 36 months of observations. This procedure results in 7,806 t-statistic observations. Table 2.2 shows the maximum-likelihood estimates for the short-run sample. Overall the distribution of short-run skill is remarkably similar to the distribution of long-run skill. The primary difference is that the means of the poor and good performers are slightly more extreme in the short-run than in the long-run. These results match our intuition that extreme short-run performances

Table 2.2: Parameter estimates based on the EM algorithm

	Full Sample: 1975-2015					
	Long-Run			Short-Run		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.3301	-0.5496	1.0760	-2.6396	-0.3740	1.6500
	(0.3044)	(0.0708)	(0.1228)	(0.0922)	(0.0368)	(0.2832)
σ_j	0.8499	1.0475	0.8518	1.0246	1.1995	1.1360
	(0.1290)	(0.0317)	(0.0574)	(0.0380)	(0.0194)	(0.1196)
π_j	0.0778	0.8379	0.0844	0.0908	0.8385	0.0707
	(0.0174)	(0.0139)	(0.0245)	(0.0088)	(0.0061)	(0.0104)
	Barras et al. Sample: 1975-2007					
	Long-Run			Short-Run		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.6449	-0.6036	1.2372	-2.6977	-0.3877	1.9183
	(0.3832)	(0.0728)	(0.2411)	(0.3758)	(0.0605)	(0.5631)
σ_j	0.8532	1.0510	0.8212	0.9219	1.1526	1.0289
	(0.1696)	(0.0445)	(0.0985)	(0.1606)	(0.0588)	(0.2457)
π_j	0.0841	0.8392	0.0767	0.0844	0.8390	0.0766
	(0.0261)	(0.0198)	(0.0196)	(0.0237)	(0.0398)	(0.0268)

Note: The table shows the parameter estimates for the distribution of mutual fund t-statistics.

are less likely over the long-run.

2.4.3 The evolution of skill

Having seen the evidence in favor of a large population of good performers, I turn next to an investigation of how the distribution of fund skill has evolved. To do so, I perform the following experiment: for each year I repeat the long-run estimation procedure using the entire time series of returns up to the beginning of that year, starting with 1990 and ending with 2015. This process results in times series for the parameter estimates $\{\mu_t, \sigma_t, \rho_t\}_{t=1990,\ldots,2015}$.

The results are shown in Figure 2.2. The time series of estimated means do not demonstrate any strong trends over the sample period. On the contrary, the estimates have been remarkably stable, particularly over the last ten years. The time series of estimated weights also display little variation over the last decade.

These results are much different from the time series results in Barras et al. (2010). In that paper, the authors find a strong downward trend in the size of the skilled population, as well as a corresponding increase in the population of unskilled funds. While a reduction in the size of the skilled manager population is consistent with the theoretical predictions found in Berk and Green (2004) and Pastor and Stambaugh (2012), a dramatic increase in the size of the unskilled population requires further explanation.

2.4.4 The persistence of skill

The results thus far demonstrate that fund returns are consistent with the presence of multiple levels of manager ability and that the properties of these classes are stable over time. It is possible, however, that the constituents of the subpopulations are unstable, making it difficult to capitalize on the performance of the best funds. To address this issue, I use a classification scheme to construct portfolios of above and below average performing funds and evaluate their out-of-sample performance.

Starting from January 1, 1980, each year I estimate the skill distribution parameters using the preceding five years of returns. I include all funds that have at least 36 months of observations over this period. Using these estimates, I calculate

the conditional probability that each return belongs to the above average population given its estimated t-statistic. I then construct five equal-weighted portfolios of good performers based on fixed probability thresholds $\tau = 0.50, 0.60, \ldots, 0.90$. The classification rule is simple: $P(z_i = j; \hat{\Theta}) > \tau \Rightarrow \hat{z}_i = j$. If no good funds are found at a particular threshold, then the returns over the holding period for that portfolio are the monthly market returns ($\alpha = 0$). Otherwise, the portfolio is held for one year, after which I repeat the classification procedure. For comparison, I form corresponding portfolios of poor performing funds, as well as a portfolio of "average" funds.

Table 2.3 describes the composition of the resulting portfolios using the proportion of funds assigned to each portfolio. The proportion of funds in each portfolio is ordinarily smaller than the estimated proportion of funds belonging to the corresponding class due to the threshold criteria. In fact, in many years no funds are assigned to the portfolios constructed using the highest threshold: 13 (14) out of 35 years I fail to find any good (poor) funds with high probability.

Table 2.4 shows the turnover of each portfolio. For each portfolio, I calculate the probability that a fund assigned to that portfolio is re-assigned to that portfolio in future years. The results show that there is moderate persistence over short time horizons of up to two years and that the level of persistence decreases as the threshold increases. For example, at the 80% threshold level, I find that there is a 20% chance that a fund is re-assigned to either of the good or poor classes one year after being assigned to the same portfolio. At the 90% threshold level, the probability drops to less than 10%.

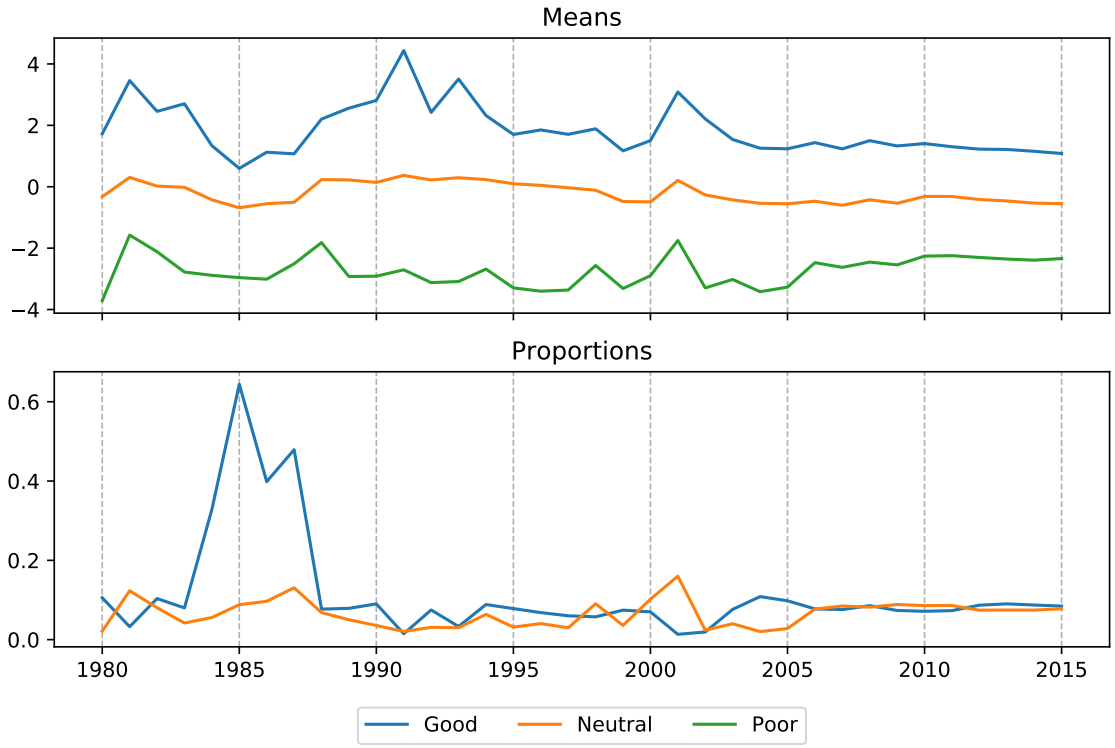
Table 2.3: Fund classification proportions

			Proportion				
Population	τ	$\bar{\pi}_c$	$= 0\%$	$0 - 6\%$	$6 - 12\%$	$12 - 24\%$	$> 24\%$
Poor	0.90	0.0257	13	21	0	0	1
Poor	0.80	0.0365	6	26	2	0	1
Poor	0.70	0.0496	2	27	5	0	1
Poor	0.60	0.0629	0	28	4	2	1
Poor	0.50	0.0787	0	21	10	2	2
Ave.	0.00	0.8381	0	0	0	2	33
Good	0.50	0.0833	0	23	8	2	2
Good	0.60	0.0695	0	27	5	1	2
Good	0.70	0.0585	3	26	3	1	2
Good	0.80	0.0469	8	23	1	1	2
Good	0.90	0.0348	14	18	1	1	1

Note: At the beginning of each formation year every fund with at least 36 months of returns during the preceding 60 months is classified according to its estimated conditional probability of belonging to either the *Bad*, *Neutral*, or *Good* population of funds. For each population an equal-weighted portfolio is formed and held for the following 12 months. This process is repeated at the beginning of each year from 1980 to 2014. If a fund disappears during a holding period, then it is dropped and its weight reassigned to the remaining funds. A fund is classified as *Neutral* if it is not assigned to any of the *Bad* or *Good* populations with the probabilities shown. $\hat{\pi}_{pop}$ is the average proportion of funds assigned to the corresponding population. The remaining values are the number of years for which the estimated proportion of funds in the corresponding population ($\hat{\pi}_t$) falls in the corresponding interval.

Table 2.5 analyzes the performance of each portfolio. I evaluate the performance of portfolios using the same four-factor model used to calculate individual fund t-statistics. The results show that portfolios of the best funds generate positive out-of-sample alphas, although none of these are statistically significant by conventional standards. Contrary to what one might expect, the alphas are not increasing (decreasing) in τ for the good (poor) portfolios. The reasons for this are that the

Figure 2.2: Time series of parameter estimates



Note: (Top) For year from 1980 to 2015 I plot the time series of the estimated populations means of the mixture model for t-statistics using fund returns up to that year. (Bottom) The estimated populations weights of the mixture model for t-statistics using fund returns up to that year.

classification is imperfect, and the default passive investment is more likely to be used at higher thresholds. Also, note that the performances of the poor class are highly statistically significant, as is the performance of the average portfolio. Thus our classification procedure provides a substantial benefit to an investor choosing amongst actively managed funds.

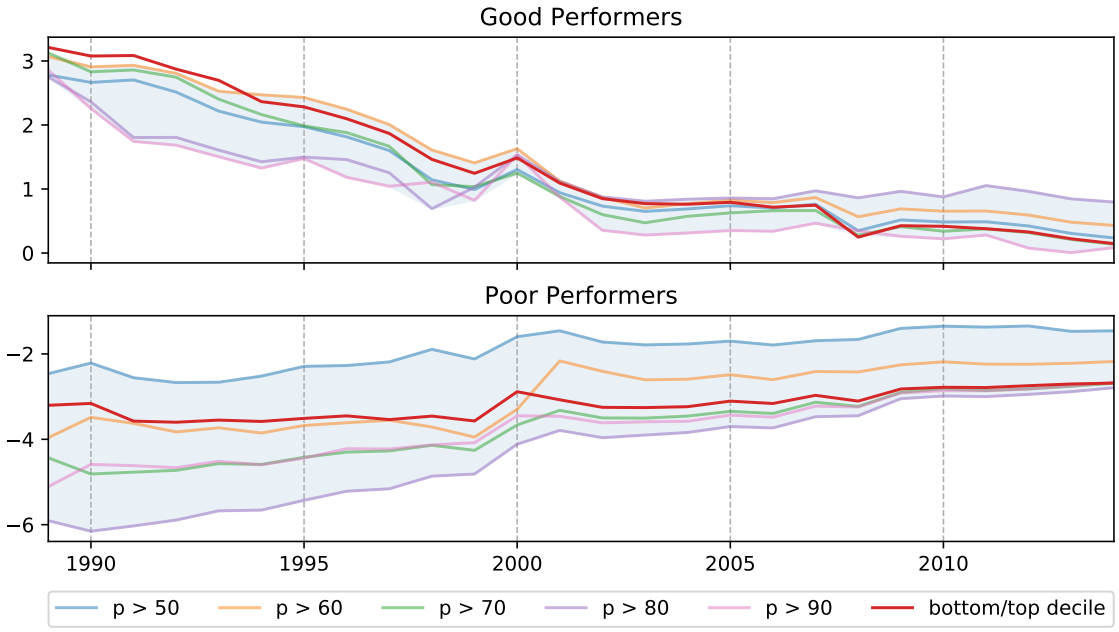
Figure 2.3 shows the performance of the good and poor portfolios over time.

From the time series of monthly portfolio returns, I construct time series of portfolio alphas by running four-factor regressions at the beginning of each year, starting in January 1990. The results for the top performers tell a different story from Figure 2.2: the portfolio alphas decrease steadily, suggesting that either the size or ability of the best performing population declined over this part of the sample. The disparity may be due to the fact that the earlier results use expanding windows, but it also a reflects the high turnover documented in Table 2.4.

Care is required when interpreting the results shown in Figure 2.3. What the figure shows is that before 1990 the average performance of the best portfolio was excellent. The fact that the time series is decreasing over the sample period shown indicates that the portfolio alphas over this period are at best unimpressive, and perhaps negative. When we compare the time series of the classified portfolios with the performance of the naive portfolio, the fact that the alpha of the naive portfolio decreases more indicates that it performs particularly bad over this range. The reason is that naive portfolio does not account for changes in the distribution of skill. On the other hand, portfolios based on the classification method automatically adjust to such changes and shift from active to passive strategies.

Based on the results in Table 2.4 we know that the composition of portfolios changes frequently and dramatically. Instead of measuring performance by alpha as in Table 2.5, which assumes that the factor loadings of the portfolios do not vary over time, we can look at errors based on continually evolving measurements of factor loadings. Specifically, each year after forming portfolios I calculate the factor loadings

Figure 2.3: Portfolio performance



Note: (Top) Time series alpha estimates for the portfolios of good performers. (Bottom) Time series alpha estimates for the portfolios of poor performers.

from a four-factor model, $\mathbf{B}_t$, and compute the monthly alphas for each of the next twelve months using the formula $\alpha_{t+i} = r_{t+i} - \mathbf{B}_t \mathbf{f}_{t+i}$ $(i = 1, 2, \ldots, 12)$. This procedure results in a time series of monthly alphas. The results of this approach are displayed in Table 2.5. I find that the average annualized monthly alphas of portfolios of poor and average performers are substantially lower than those of portfolios of good performers and that all portfolios of good performers have positive average alphas. The results agree with Table 2.5: despite high fund turnover, the average performance of the classified portfolios agrees with their type and the estimates in Table 2.2.

2.4.5 Estimate validation

Expectation-maximization produces numerical approximations to maximum likelihood estimates. Theoretically, maximum likelihood estimates have desirable asymptotic properties: most importantly, they unbiased and efficient. However, the expectation-maximization algorithm might struggle to learn the parameters of a mixture of normals model if some of the weights are particularly small, or if the densities of the individual distributions have a large amount of overlap (i.e. if the means of the distributions are close relative to their standard deviations). In either case, the algorithm is more likely to converge to a local maximizer. To validate the estimation procedure, I perform a simple experiment in which I estimate the parameters of mixture of normal distributions using simulated data.

For a fixed set of parameters, I generate 1000 samples containing 3000 pseudo-random observations each. Using the same expectation-maximization algorithm applied above, I estimate the parameters of each sample. I then investigate the finite sample distribution of the resulting 1000 estimates.

Table 2.6 shows the results. I compare two choices of parameters, one based on inferred values from Barras et al. (2010), and the other based on my estimates. Most of the parameters are unbiased: the exceptions are the mean and weight of the good performers under the distribution based on Barras et al. (2010). In this case, the population of good funds is simply too small to identify—in fact, the proportion estimated by the authors is not statistically different from zero.

These results present a dilemma: If the mixture of normals model is correct,

then the estimated weights should be similar to those found in Barras et al. (2010). In Table 2.2, Panel B I show that using the same sample period as in that paper, this is not the case. A possible explanation for the discrepancy is that the estimation procedures are sensitive to the treatment of the average population's mean. We may find more similar results if we restrict the mean of this class to zero.

2.5 Conclusion

In this paper, I investigated the distribution of skill in the actively-managed equity mutual fund industry. My analysis is based on a simple framework for which precise estimation methods are available. I demonstrate how to apply this structure to evaluate skill in the industry as a whole, and also perform fund-level evaluations. My results show that over an extended sample period the distribution of skill is consistent with the existence of substantial populations of both poor and good performing funds, as well as a majority population of managers that under-performs low-cost, passive investment alternatives. I find mixed results concerning the evolution of the distribution of skill. My formal estimates indicate that this distribution has remained stable over time, but the performance of an investment strategy based on these estimates suggests a decline in either the skill or proportion of skilled funds in the distribution over the latter half of our sample period. I validate these results using a simple simulation exercise.

What are we to conclude of the active management industry? On the one hand, it appears that there is a small number alpha generating managers to be found

in the distribution of funds at any particular time. At the same time, predicting which funds will perform well in the future is difficult because the size of the talented pool is small, and its expected excess return is not much larger than that of the average fund. For an investor in this asset class, the easier task is avoiding under-performing funds.

Table 2.4: Fund classification turnover

		Years After Classification				
Population	τ	1	2	3	4	5
Poor	0.90	0.0578	0.0588	0.0441	0.0195	0.0129
Poor	0.80	0.1962	0.1149	0.0595	0.0372	0.0329
Poor	0.70	0.2651	0.1621	0.1005	0.0681	0.0491
Poor	0.60	0.3272	0.2032	0.1292	0.0877	0.0564
Poor	0.50	0.3684	0.2435	0.1538	0.1065	0.0689
Ave.	-	0.8993	0.8539	0.8136	0.7831	0.7509
Good	0.50	0.3428	0.1916	0.1002	0.0754	0.0419
Good	0.60	0.2939	0.1613	0.0756	0.0609	0.0352
Good	0.70	0.2390	0.1312	0.0526	0.0540	0.0324
Good	0.80	0.1930	0.1169	0.0372	0.0525	0.0263
Good	0.90	0.0909	0.1010	0.0209	0.0383	0.0105

Note: At the beginning of each formation year every fund with at least 36 months of returns during the preceding 60 months is classified according to its estimated conditional probability of belonging to either the *Bad*, *Neutral*, or *Good* population of funds. For each population an equal-weighted portfolio is formed and held for the following 12 months. This process is repeated at the beginning of each year from 1980 to 2014. If a fund disappears during a holding period, then it is dropped and its weight reassigned to the remaining funds. A fund is classified as *Neutral* if it is not assigned to any of the *Bad* or *Good* populations with the probabilities shown. $\hat{\alpha}$ is the intercept from the regression $\tilde{r}_{p,t} = \alpha + \beta^\top f_t$. p-values are calculated from t-statistics based on homoskedastic standard errors. $\hat{\alpha}$, $\bar{r}$, and σ are annualized.

Table 2.5: Fund performance by classification

Population	τ	$\hat{\alpha}$	p	$\hat{b}_{mkt}$	$\hat{b}_{smb}$	$\hat{b}_{hml}$	$\hat{b}_{umd}$	$\bar{r}_p$	σ_p
Poor	0.90	-2.0042	0.0120	1.0053	0.1036	0.0304	0.0394	6.3884	1.6443
Poor	0.80	-2.1981	0.0023	1.0183	0.1566	0.0226	0.0340	6.3103	1.6733
Poor	0.70	-2.7453	0.0004	0.9814	0.1321	0.0447	0.0258	5.4578	1.6117
Poor	0.60	-3.0279	0.0002	0.9931	0.1464	0.0391	0.0331	5.3219	1.6398
Poor	0.50	-2.7651	0.0000	0.9974	0.1571	0.0325	0.0193	5.5094	1.6339
Ave.	-	-1.0887	0.0070	0.9930	0.2128	-0.0100	0.0164	7.0635	1.6359
Good	0.50	0.2181	0.6689	0.9547	0.3339	-0.0444	0.0038	8.0436	1.6549
Good	0.60	0.5612	0.2739	0.9428	0.3317	-0.0409	-0.0102	8.1998	1.6368
Good	0.70	0.4212	0.4887	0.9561	0.3317	-0.0534	-0.0134	8.0944	1.6727
Good	0.80	0.7060	0.2833	0.9560	0.2676	-0.0276	-0.0323	8.2341	1.6435
Good	0.90	0.5404	0.4887	0.9534	0.2027	-0.0006	-0.0299	8.0620	1.6205

Note: At the beginning of each formation year every fund with at least 36 months of returns during the preceding 60 months is classified according to its estimated conditional probability of belonging to either the *Bad*, *Neutral*, or *Good* population of funds. For each population an equal-weighted portfolio is formed and held for the following 12 months. This process is repeated at the beginning of each year from 1980 to 2014. If a fund disappears during a holding period, then it is dropped and its weight reassigned to the remaining funds. A fund is classified as *Neutral* if it is not assigned to any of the *Bad* or *Good* populations with the probabilities shown. The values in this table are estimates of the probability that a fund assigned to a particular population at time t is assigned to the same population at time $t + s$ for $s = 1, \ldots, 5$.

Table 2.5 (continued)

Population	τ	$\bar{\alpha}_t$	$\sigma(\alpha_t)$	$q_{0.05}$	$q_{0.10}$	$q_{0.50}$	$q_{0.90}$	$q_{0.95}$
Poor	0.90	-1.8976	0.3782	-22.4486	-14.6903	0.0000	9.0749	14.9710
Poor	0.80	-2.8341	0.3631	-20.0108	-14.8880	-0.7921	7.7922	13.0952
Poor	0.70	-3.8121	0.4241	-28.1429	-17.5550	-2.4551	8.1537	14.4757
Poor	0.60	-4.1412	0.4389	-29.5569	-19.9815	-3.2508	8.7368	15.8375
Poor	0.50	-3.4729	0.3715	-23.9097	-16.3740	-2.3868	8.7435	15.0773
Ave.	-	-1.0487	0.2534	-13.0754	-10.1157	-1.3581	7.9048	11.2645
Good	0.50	0.2762	0.3364	-15.5591	-11.4703	0.4807	11.1805	16.2520
Good	0.60	0.4129	0.3394	-15.5113	-11.6139	0.7028	11.8096	17.4997
Good	0.70	0.5706	0.3813	-16.9521	-12.6671	0.0000	13.9007	17.7370
Good	0.80	0.5232	0.3369	-14.6665	-11.6031	0.0000	12.6060	17.7770
Good	0.90	0.7665	0.3500	-17.4433	-12.0193	0.0000	13.9968	20.7268

Note: At the beginning of each formation year every fund with at least 36 months of returns during the preceding 60 months is classified according to its estimated conditional probability of belonging to either the *Bad*, *Neutral*, or *Good* population of funds. For each population an equal-weighted portfolio is formed and held for the following 12 months. This process is repeated at the beginning of each year from 1980 to 2014. If a fund disappears during a holding period, then it is dropped and its weight reassigned to the remaining funds. A fund is classified as *Neutral* if it is not assigned to any of the *Bad* or *Good* populations with the probabilities shown. Monthly portfolio alphas (α_t) are calculated as $r_t - \hat{\beta}_t^\top \mathbf{f}_t$, where $\hat{\beta}_t^\top$ is estimated from the 60 months of returns preceding the formation date. $\bar{\alpha}_t$ and $\sigma(\alpha_t)$ are the time series mean and standard deviation of α_t. All values are annualized.

Table 2.6: Monte Carlo simulation

	Panel A: Barras et al. (2010)					
	Estimate			Bias		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.5109	-0.0213	2.2712	-0.0109	-0.0213	-0.7288
σ_j	0.9873	0.9877	1.2333	-0.0127	-0.0123	0.2333
π_j	0.2293	0.7332	0.0375	-0.0007	-0.0168	0.0175
	Panel B: Expectation-Maximization					
	Estimate			Bias		
	Poor	Average	Good	Poor	Average	Good
μ_j	-2.3764	-0.5435	1.1492	-0.0264	0.0065	0.0692
σ_j	0.8185	1.0346	0.7995	-0.0315	-0.0154	-0.0505
π_j	0.0844	0.8363	0.0793	0.0044	-0.0037	-0.0007

Note: The reported estimates are the mean of 1000 estimates of the mixture of normal distribution parameters. Each estimate is based on 3000 pseudo-random observations.

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)

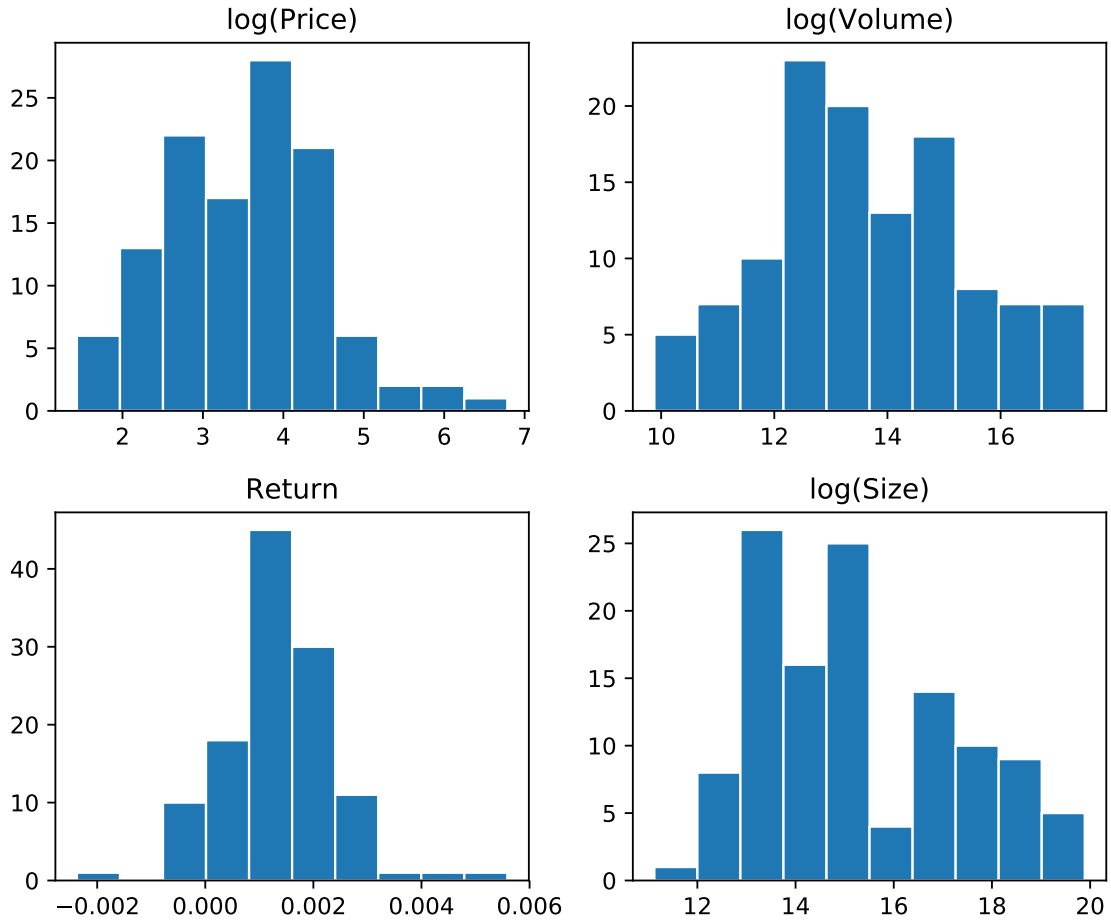
The table shows the results of regressing the estimated vector autoregression coefficients on high-frequency trading: $Y_{i,t} = \alpha + \beta_1 HFT + \beta_2 \log(Price) + \beta_3 \log(Volume) + \beta_4 Ret_{i,t} + \beta_5 \log(Size) + \epsilon_{i,t}$. HFT is the the logarithm of the daily number of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

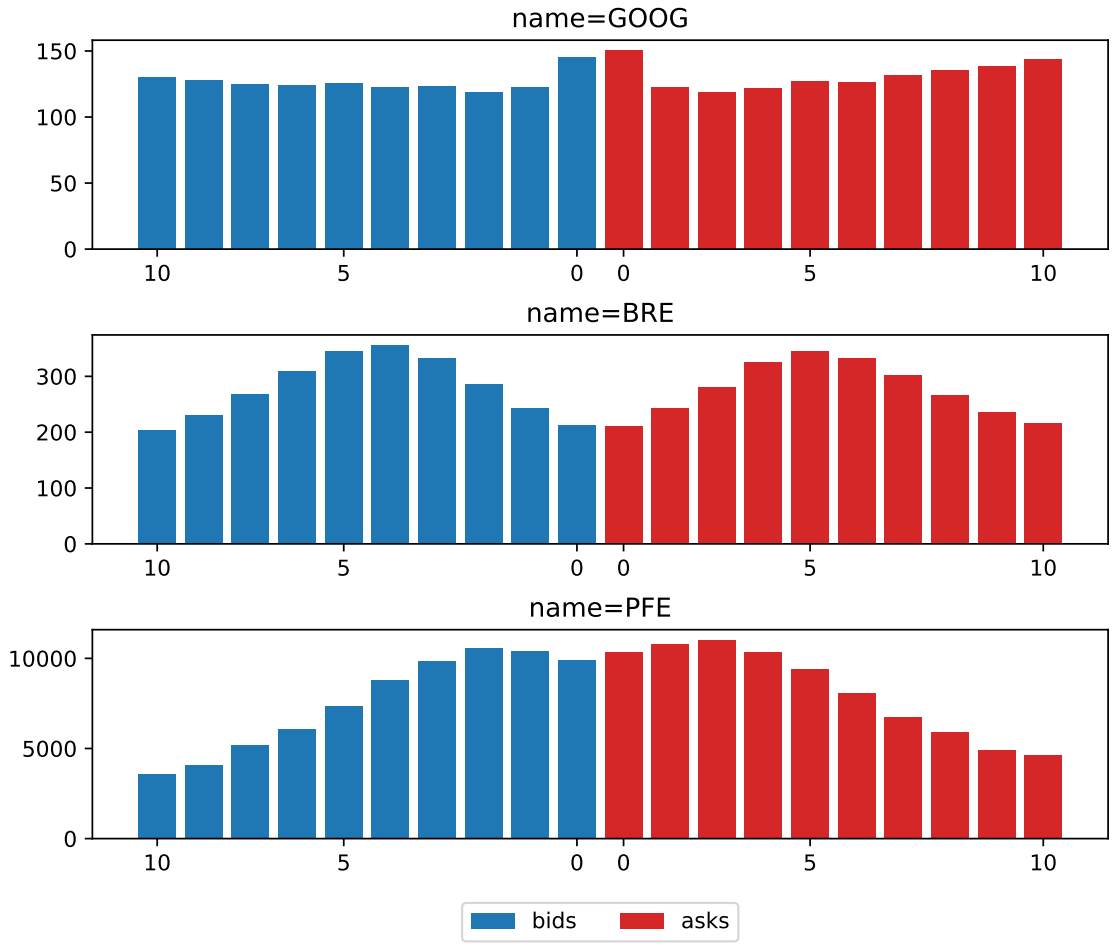
The table shows the results of regressing the estimated vector autoregression coefficients on high-frequency trading: $Y_{i,t} = \alpha + \beta_1 HFT + \beta_2 \log(Price) + \beta_3 \log(Volume) + \beta_4 Ret_{i,t} + \beta_5 \log(Size) + \epsilon_{i,t}$. HFT is the the logarithm of the daily number of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



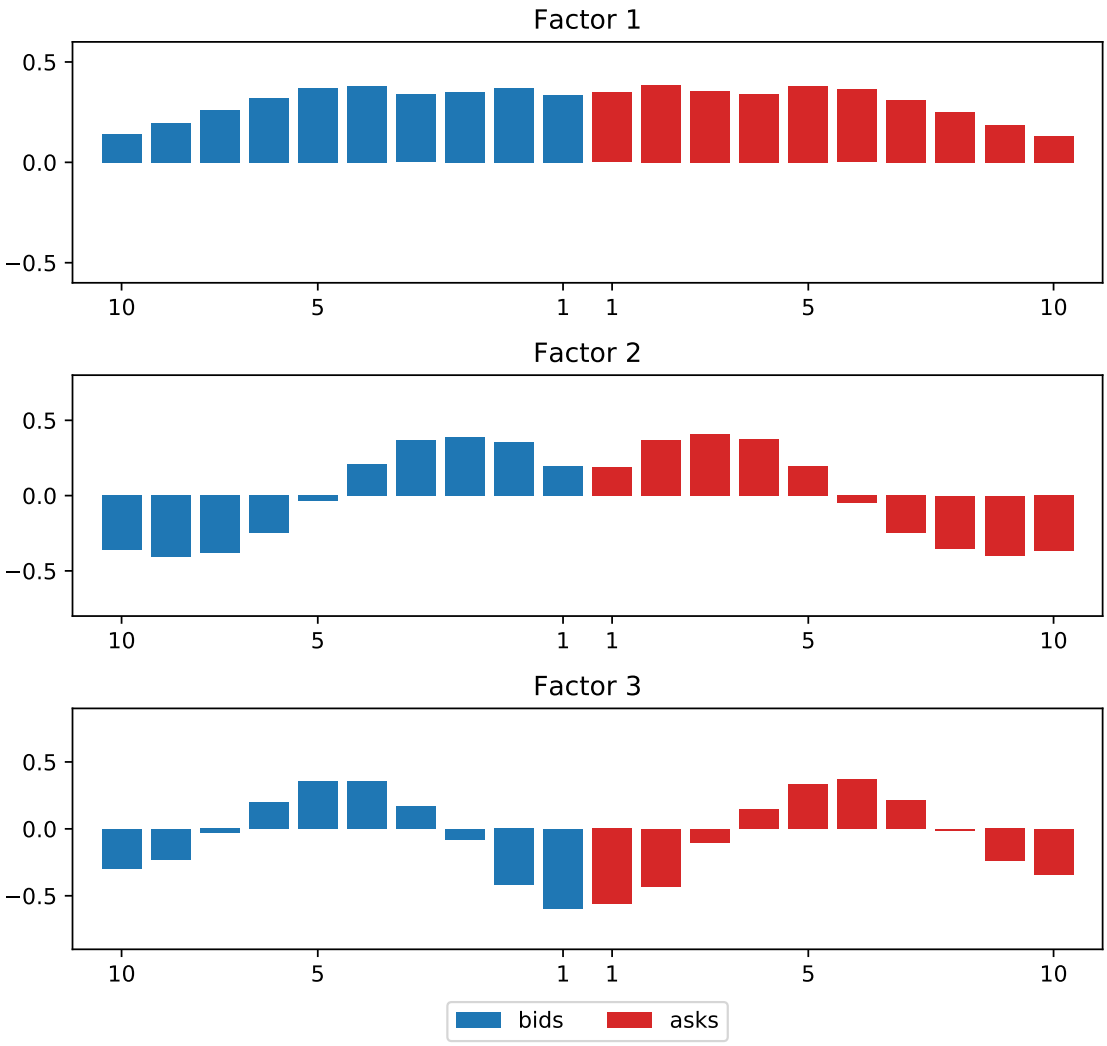
Note: These figures demonstrate the distribution of daily average stock characteristics for my sample. $\log(Price)$ is the logarithm of the daily closing price; $\log(Volume)$ is the logarithm of the daily number of shares sold; $Return$ is the daily holding period return in basis points; $\log(Size)$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



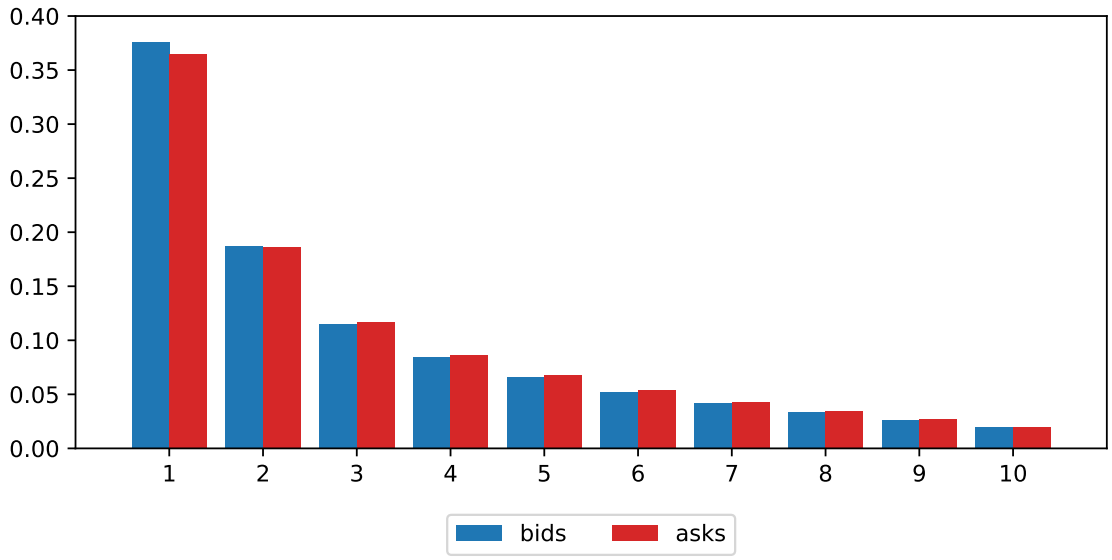
Note: The figure shows the average shape of randomly selected large (top row), medium (middle row), and small (bottom row) firms. Average shapes are calculated from all order book updates in the ITCH database during 2013.

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.

Figure 3.4: Principal component analysis of order books (cont.)



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

4.1 Introduction

4.1.1 Summary

4.1.2 Contribution

4.2 Methodology

4.2.1 Model

Order book dynamics are difficult to understand, but it is useful to think of two aspects of the trading environment: an external (endogenous) environment and an internal (endogenous) environment. Some orders are motivated by considerations that are external to the market as well as the information about the asset. Other orders are motivated by considerations based on external information about the asset. Still other orders are placed by traders that are indifferent to the external information?their orders only reflect the state of the market itself. The motivations are reflected in market microstructure models as liquidity traders, informed traders and market makers.

Limit order books are event-driven processes. In order to come up with a model that can serve as a simulator, we need a statistical framework that works in continuous-time; we need to be able to distinguish the order of events exactly. In addition, sequences of events generated by the simulator should reflect the actions of the user. In other words, the user should be able to change the expected outcome

of the simulation, rather than simply learning the expectation. Point processes–and Poisson processes in particular–are statistical models that satifsy both criteria.

4.2.1.1 Poisson Processes

A *Poisson process* is a continuous-time model that simulates a sequence of events $\{s_m\}_{m=1}^{M}$ occurring within a time interval $[0, T]$. The expected number of events that occur in any sub-interval $[\tau_1, \tau_2]$ is determined by the *intensity* of the process, $\lambda(t)$. If the intensity is constant, then the expected number of events is $\lambda(\tau_2 - \tau_1)$. In general, the intensity is time-varying (e.g., in order to capture seasonalities in the frequency of events), in which case the expected number of events is equal to the *integrated intensity*,

$$\mathcal{I}[t_1, t_2] = \int_{\tau_1}^{\tau_2} \lambda(t)\ dt = \mathbb{E}[N] \tag{4.1}$$

A *self-exciting Poisson process*–or *Hawkes process* (Hawkes (1971))–is a type of inhomogeneous Poisson process. At any point in time, the intensity of a self-exciting Poisson process depends on the history of events preceding that point. This self-exciting feature allows the model to capture temporary increases in the probability of future events. For example, gang-related crimes increase the probability of subsequent gang-related crimes because they are followed by a cycle of revenge crimes (Cho et al. (2013)).

The intensity of a self-exciting Poisson process is given by

$$\lambda(t) = \lambda^0(t) + \sum_{s_m < t} \phi(t - s_m), \tag{4.2}$$

where ϕ is a non-negative *impulse-response* function that produces an increase in the

occurrence of events following an initial event. In the absence of events, the model functions as a standard Poisson process. When an event occurs, the model generates a momentary burst in the intensity of the event arrivals. If the process is *stable*, it produces clusters of events. The process is *unstable* if the feedback loop between events causes the intensity to explode.

Self-exciting Poisson processes can be simulated using the *Poisson superposition principle* (Kingman (1993)), which states that a collection of (independent) Poisson processes is equivalent to an individual Poisson process. More precisely, suppose that we are shown the event data from a collection of K Poisson processes, $\{\{s_m^{(k)}\}_{m=1}^{M_k}\}_{k=1}^{K}$, but that the data have been merged into a single sequence of events without indicating their origins. Thus, we observe $\{s_m\}_{m=1}^{\sum_k M_k}$, and there is no difference, from our perspective, between assuming that the data is generated by a collection of processes with intensities $\{\lambda_k(t)\}_{k=1}^{K}$ and assuming that it is generated by a solitary Poisson process with intensity $\lambda_{tot}(t) = \sum_{k=1}^{K} \lambda_k(t)$.

For a self-exciting Poisson process, the superposition principle implies a generative model in which each event is the offspring of either the "background" Poisson process, or of a Poisson process initiated by an earlier event. Therefore, the process can be simulated according to the following algorithm:

1. Generate a sequence of events according to a standard Poisson process with intensity $\lambda_0(t)$ on the interval $[0, T]$.
2. For each event $s_m \in [0, T]$, generate a sequence of events according to a Poisson process with intensity $\phi(t - s_m)$ on the interval $[s_m, T]$.

3. Repeat Step 2 until no further events are occur.

4.2.1.2 Poisson Processes on Networks

Self-exciting Poisson processes are easily extended to multiple processes. The multivariate model allows for connections between different types of events in addition to connections between events of the same type. Specifically, the intensity of the n^{th} type of event is give by

$$\lambda_n(t) = \lambda_{n,0}(t) + \sum_{s_m<t} h_{c_m,n}(t - s_m; \theta_{c_m,n}) \tag{4.3}$$

The generative model relies on a clever trick explained in Linderman (2016). Suppose that we had a collection of Poisson processes $\{\{s_{m_i}\}_{m_i=1}^{M_i} \sim \mathcal{PP}(\lambda_i(t)) \mid i = 1, \ldots, K\}$. The superposition principle asserts that the union of this collection is itself a Poisson process with parameter $\lambda(t) = \sum_{i=1}^{N} \lambda_i(t)$. The reverse is also true: if we know that the parameter of a Poisson process has the form $\lambda(t) = \sum_{i=1}^{N} \lambda_i(t)$, then the process is equivalent to a union of Poisson processes with parameters $\lambda_i(t)$. The functional form of the intensity parameter for the Hawkes process in equation (4.3) has exactly this form. Moreover, there is a one-to-one correspondence between the collection of Poisson processes implied by equation (4.3) and the events of the original process. In particular, each event can be interpreted as a Poisson process with intensity function

$$\lambda_n^m(t) = \begin{cases} 0 & t \leq s_m \\ h(t - s_m; \theta_{c_m,n}) & t > s_m. \end{cases} \tag{4.4}$$

(A figure that demonstrates a multivariate Hawkes process).

4.2.1.3 Order Book Model

Poisson networks provide a *general* framework for modeling limit order book dynamics, but precisely how to use this framework depends on what aspect of limit order books we are interested in. We could use this framework to analyze the connections between limit order books: How does activity on order book A affect activity on order book B? Our objective is instead to construct a simulator of an individual limit order book without reference to external forces. We are interested in capturing the connection between different *types* of orders and between orders affecting different *levels* of the order book. Therefore the nodes of our model are pairs of *type-level* combinations. The types of orders that we consider are *add*, *delete*, and *execute*. We are only interested in the first three levels of the order book, and so this results a graph with dimension $K = 3 \times 3 = 9$.

The dynamics of a limit order book are an event-driven process. Over longer periods of time the effects of individual events are smoothed out and it is reasonable, and practical, to model the book in terms of "residual" metrics: prices and shares. Over short time periods we need to model the book at the level of individual events, which is our motivation for using Poisson processes. But what are the classes of events that we should be interested in: what are the nodes of our graph? All orders can be described by a small set of properties. For example, in its out-going message data the NASDAQ exchange describes orders in terms of order type (add, delete, execute), side (bid, ask), price, number of shares, and a unique order reference number (used to match delete and execute orders with add orders). Certainly we should draw

distinctions between order types and orders on different sides of the book. Prices are problematic because they can have a wide range of values, which would require us to create a large graph. More importantly, what is the connection between a bid add order at $98.01 when the best bid is $98.02, and a bid add order at $98.01 when the best bid is $98.10? Prices don't define meaningful classes of events, but the *levels* of the order book that prices correspond to do. I define the level of an order to be the number of cents away from the best bid/ask that the order price corresponds to. So an add order at the current best bid price has level 0, and an add order one cent below the current best bid price has level 1. Add orders can also specify prices that improve upon the current best bid/ask, which we label as level -1 orders; delete and execute orders can never have level -1 (what about execution of "iceberg" orders?). Sometimes a market order executes against multiple levels of the order book, which could be considered as two events occurring at the same time: an execute at the first level, and an execute at the second level. In order to account for this possibility we can define the level of an execute order as the maximum of the levels affected by the order. Execute orders are almost always level 1, but will be classified as level 2 when an order "walks the book"[1]. In order to keep the dimension of the graph tractable, we only consider add order events at levels -1, 0, and 1, an innocuous assumption given that orders outside of the top of the book don't play an important role in order book dynamics. Overall, this leaves us with 14 classes of events:

---

[1]Report how often this occurs in your dataset.

$$k = \sum_{\substack{s\in\{bid,ask\},\\ l\in\{-1,0,1\}}} \mathbb{I}(type = add, side = s, level = l) + \sum_{\substack{s\in\{bid,ask\},\\ l\in\{0,1\}}} \mathbb{I}(type = delete, side = s, level = l) + \sum_{\substack{s\in\{bid,ask\},\\ l\in\{0,1\}}} \mathbb{I}(type = execute, side = s, level = l). \quad (4.5)$$

In some cases I find that *no* market orders execute against the entire available volume at the best bid/ask, in which case I reduce the size of the graph by a further 2 classes.

The Poisson network model allows for a variety of dependencies between network nodes. A *fully dependent* order model is a Poisson network with a fully-connected graph: the binary connection matrix uniformly one ($A = 1$). In a *Bernoulli* order model, some nodes are connected, and some nodes are independent: whether two nodes are connected is determined by the flip of a coin, and the same coin is used for each pair of nodes in the network (we could consider more complicated strucutres in which different coins are used for each pair of nodes). In our model of the limit order book we assume that every node is connected to every other node. Assuming that the network is fully-connected simplifies inference as we only need to worry about the strength between nodes (W), which reduces the computational complexity of the Gibbs sampling algorithm.

4.2.2 Inference

Estimation of Hawkes processes in the finance literature relies on maximum-likelihood methods (e.g., Bowsher (2007), Large (2007), and Bacry et al. (2013)).

I introduce a Bayesian inference procedure to the finance literature. The method is based on work in the field of computational neuroscience, first published in Linderman (2016).

4.2.2.1 Modelling Choices

The model framework is the network Poisson process model from the previous section. We assume that $\lambda_0(t) = \lambda$.

(Write down assumptions about the impulse-response functional form. Why this form? What types of shapes can it take?)

4.2.2.2 Gibbs Sampling Algorithm

We can sample the conditional posterior distributions of the model parameters using conjugate priors and Markov Chain Monte Carlo methods .

(*Weights*) If the weights have gamma priors,

$$w_{n,n'} \sim \Gamma(\kappa, \nu_{n,n'}), \tag{4.6}$$

then their conditional distributions are also gamma:

$$p(w_{n,n'}|\{s_m, c_m, \omega_m\}_{m=1}^M, a_{n,n'} = 1, \kappa, \nu_{n,n'}) = \Gamma(\tilde{\kappa}_{n,n'}, \tilde{\nu}_{n,n'}), \tag{4.7}$$

where

$$\tilde{\kappa}_{n,n'} = \kappa + M_{n,n'}, \tag{4.8}$$

$$\tilde{\nu}_{n,n'} = \nu_{n,n'} + M_n, \tag{4.9}$$

and

$$M_n = \sum_{m=1}^{M} \mathbb{I}\left[c_m = n\right], \tag{4.10}$$

$$M_{n,n'} = \sum_{m=1}^{M} \mathbb{I}\left[c_m = n\right] \mathbb{I}\left[c_{m'} = n'\right] \mathbb{I}\left[\omega_{m'} = m\right]. \tag{4.11}$$

The first sufficient statistic (M_n) is the number events occuring on node n. The second sufficient statistics ($M_{n,n'}$) is the number of events on node n' caused by events on node n.

(*Impulses*) The likelihood is conjugate with a normal-gamma distribution. If the priors of the impulse parameters are normal-gamma,

$$(\mu_{n,n'}, \tau_{n,n'}) \sim \mathcal{NG}(\mu_\mu, \kappa_\mu, \alpha_\tau, \beta_\tau), \tag{4.12}$$

then the conditional posterior distributions are normal-gamma with parameters

$$\tilde{\mu}_{n,n'} = \frac{\kappa_\mu \mu_\mu + M_{n,n'} \bar{x}_{n,n'}}{\kappa_\mu + M_{n,n'}}, \tag{4.13}$$

$$\tilde{\kappa}_{n,n'} = \kappa_\mu + M_{n,n'}, \tag{4.14}$$

$$\tilde{\alpha}_{n,n'} = \alpha_\tau + \frac{M_{n,n'}}{2}, \tag{4.15}$$

$$\tilde{\beta}_{n,n'} = \frac{\nu_{n,n'}}{2} + \frac{M_{n,n'} \kappa_\mu (\bar{x}_{n,n'} - \mu_\mu)^2}{2(M_{n,n'} + \kappa_\mu)}, \tag{4.16}$$

The sufficients statistics in this case are defined with respect to

$$x_{m,m'} = \ln\left(\frac{s_{m'} - s_m}{\Delta t_{max} - (s_{m'} - s_m)}\right), \tag{4.17}$$

which is the log of the ratio of the time elapsed since event m occured to the time remaining until event m can no longer cause another event. The sufficient statistics

are the mean and variance of $x_{m,m'}$:

$$\bar{x}_{n,n'} = \frac{1}{M_{n,n'}} \sum_m \sum_{m'} \mathbb{I}\left[c_m = n\right] \mathbb{I}\left[c_{m'} = n'\right] \mathbb{I}\left[\omega_{m'} = m\right] x_{m,m'}, \tag{4.18}$$

$$\bar{\nu}_{n,n'} = \frac{1}{M_{n,n'}} \sum_m \sum_{m'} \mathbb{I}\left[c_m = n\right] \mathbb{I}\left[c_{m'} = n'\right] \mathbb{I}\left[\omega_{m'} = m\right] \left(x_{m,m'} - \bar{x}_{n,n'}\right)^2 \tag{4.19}$$

(*Biases*) If the background rates have gamma priors,

$$\lambda_n^0 \sim \Gamma(\alpha_n^0, \beta_n^0), \tag{4.20}$$

then their conditional distributions are also gamma:

$$p(\lambda_n^0 | \{s_m, c_m, \omega_m\}_{m=1}^M, \alpha_n^0, \beta_n^0) \sim \Gamma(\tilde{\alpha}_n^0, \tilde{\beta}_n^0), \tag{4.21}$$

where

$$\tilde{\alpha}_n^0 = \alpha_n^0 + \sum_{m=1}^M \mathbb{I}(c_m = n)\mathbb{I}(\omega_m = 0) = \alpha_n^0 + M_n^0, \tag{4.22}$$

$$\tilde{\beta}_0^n = \beta_0 + T. \tag{4.23}$$

In this case the sufficient statistics are the sample time (T) and the number of node n events induced by the background rate.

(*Parents*) The conditional distribution of the parent indicator variable is a multinomial with probability mass function

$$p(\omega_m = n \mid s_m, \{\lambda_n(t)\}_{n=1}^M) = \frac{\lambda_n(s_m)}{\lambda_{c_m}^0 + \sum_{n'=1}^{m-1} \lambda_{n'}(s_m)}, \tag{4.24}$$

for $n = 0, \ldots, m-1$, where

$$\lambda_{n'}(s_m) = w_{c_{n'}, c_m} \cdot \hbar(s_m - s_{n'}; \theta_{c_m, c_n'}) \tag{4.25}$$

This says that the probability that the parent of the m^{th} event is the n^{th} event equals the relative contribution of the n^{th} event to the aggregate intensity at the time of m^{th} event. As a result, the most likely parent at any time is the event that is *most active* at that time. The conditional distributions are independent, so the most likely parents can be identified in parallel. We can also see from the definition of $\lambda_{n'}(s_m)$ that the most likely parent of an event is not, in general, the most recent event: the likelihood depends on the *strength* of the connection between between the event node and the potential parent event node. In addition, events are sometimes caused by the background rate, which is reflected in the inclusion of $n = 0$ in equation (4.24).

4.3 Data

4.4 Results

What does the model reveal about limit order book dynamics in general? This section analyzes estimates of the event-driven model for a broad (by microstructure standards) cross-section of stocks. For each stock, I perform MCMC sampling for the network Poisson model with the same hyperparameter selection and settings described in previous sections using a single day of message data. The point-estimator considered is the posterior median, which minimizes posterior absolute loss risk.

All stocks belong to the S&P 500, with computational considerations determining the specific selection. I collected message data for all stocks in the S&P 500 and used this data to identify order book events according to the definitions described above. The MCMC algorithm employed scales on the order of the number of events

squared. To reduce the total sampling time, the sample excludes stocks with more than 60,000 observations between 10:30 am and 3:00 pm. An alternative approach would be to select a random subsample of manageable size from the skipped stocks. In fact, it makes more sense to use the same number of events for each stock, rather than the same amount of time, as it is this number that determines the posterior distribution, and message activity varies considerably across stocks.

The model assumes a constant background rate of events, which is an unrealistic assumption. More likely, the background rate varies over time as market conditions change. It is possible, and not computationally costly, to add a piecewise constant or linear background rate to the model. On the other hand, time-varying background rates are only generally useful if they capture seasonalities, and are complicated to analyze. As a compromise, I summarize each stock day by a single background rate parameter but restrict each day of message data to stable trading hours (10:30 am to 3:00 pm).

4.4.1 Connections

What types of interactions characterize aggregate limit order book dynamics? Figure 4.1 presents the median of the connection weight matrices. Two connections stand out immediately: add orders at the best bid following trades that increase the best ask, and add orders at the best ask following trades the decrease the best bid. This type of behavior has been noted elsewhere and shows that trading algorithms recognize these trades as dependable signals of underlying supply and demand. In

fact, just to the right of these connections are the connections between bids that exhaust the best prices and subsequent add orders that improve the offer price on the opposite side of the book. (Likely what is happening is that this connection activates first, and then new orders queue at the newly established best pricewe could confirm this by looking at the impulse responses of the two connections).

The opposite side of the connection matrix displays a related phenomenon between executions and subsequent deletions. Here, price-changing trades are seen to have two substantial effects: one on the same side of the book, and one on the opposite side. The result is the same on both sides: traders tend to delete orders at the best price. This activity is consistent with the connections between executions and additions. Traders on the opposite side of the book need to remove liquidity that will lose priority if an incoming order establishes a new price on that side. Traders on the same side of the book will cancel their orders if they anticipate that the price will continue to move in that direction (although there is no substantial connection to add orders on this side of the book at the new best price, so traders do not appear to re-establish their prior positions).

Finally, the median raw estimates demonstrate strong links between price-improving add orders and subsequent add orders at the new best prices (on the same side). Price-improving orders are significant order book events that traders respond to by supplying fresh liquidity to preserve execution priority.

All of the connections addressed thus far have been identified by looking at the raw estimates of connection weights. These raw estimates represent the expected

number of child events following parent events, so they have a direct economic meaning. Alternatively, we might ask which connections are most important relative to background activity. That is, which interactions contribute the most to the child events intensity. To do so, we simple normalize all connection weights by their child weight background rates. This normalization gives a somewhat different interpretation of the most important connections.

The pair of links that stands out the most is those between price-improving add orders and price-deteriorating delete orders. Thus, the model identifies the main driving force behind such delete orders: they are a response to price-improving orders on the same side of the book. In fact, response may not be the right description: these orders are likely to come from the same trading algorithm posted an add order above the best bid, and then immediately deleting that order if it fails to execute. (By looking at the impulse response on this connection we can get a better idea of the typical amount of time traders are willing to wait for an execution). The value of these weights is around 120, which means that the expected number of such delete orders following a price-improving add order is 120 times greater than the expected number of such delete order generated as background activity. Furthermore, the reverse link becomes clear after normalization: price-deteriorating deletions precede price-improving orders on the same side of the book. These two pairs of connections form a loop that produces cycles of flickering quotes: quick, repeated additions and deletions of price-improving liquidity.

Another pair of connections that appears much more clearly under this nor-

malization is that between executions and subsequent executions. It is well-known that trades are clustered in time. As trades are relatively rare events, this clustering is less prominent in the raw weight estimates. After normalization, we recover the observation that executions generate additional executions on the same side of the book. According to the model estimates, clustering accounts for around 60 times as many executions at the top of the book as exogenous executions at the best offer, and approximately 40 times as many for price-reducing trades.

Finally, it is interesting to note the order of importance between the execute-add connections reverses under this normalization: price-deteriorating trades have a more substantial impact on the probability of observing price-improving orders on the opposite side of the book than they do on the likelihood of add orders at the top of the book.

4.4.2 Impulse Responses

Beyond the strength of connections between events, it is necessary to discuss the timescale of these interactions. The median impulse response is the impulse response function evaluated at the posterior median point-estimator of the relevant parameters. Figure 4.2 plots the median impulse responses of the four pairs of connections discussed in the previous section.

One thing to note is that, in addition to the magnitude, the timing of the connections is symmetric in all four cases.

Figure 4.2 shows the delay in response to price-deteriorating trades of both add

orders at the opposite-side best price, and that improve the best offer on the opposite side of the book. Both reactions peak between one microsecond (1/100000) and one millisecond (1/1000), where the bulk of their mass concentrates, and of the four connections shown, they are the fastest. The impulse response of deletions to price-deteriorating executions takes place on a longer timescale, with a maximum intensity attained around one millisecond, while the intensity of delete orders following price-improving add orders peaks between one millisecond and one second. Thus, so-called "flickering orders" are in fact permitted to sit for a relatively extended period before being canceled.

It is worth briefly contrasting these results with results based on exponential impulse response functions. An exponential impulse response function creates an immediate increase in the intensity of child events. In contrast, with a logarithmic-normal impulse response, there is a lag between the time of a parent event and the time it generates a spike in the child event intensity. At timescales on the order of minutes, the distinction is unimportant. As demonstrated by Figure 4.2, however, interactions between order book events take place on timescales measured in milliseconds: on this timescale the difference is meaningful.

4.4.3 Cross-Section

The size of the sample analyzed in this study presents a unique opportunity to investigate the cross-section of order book dynamics. A question of particular interest is whether we can meaningfully categorize order books beyond the aggregate

point estimates presented above. For example, if order book dynamics determined by simple order book characteristics (e.g., the bid-ask spread), then we might expect to find distinct clusters of order books. I explore this possibility through an analysis of point estimates of the weight matrices. I observe little evidence of clusters of orders book dynamics.

Clustering refers to a form of unsupervised learning in which we attempt to find groups of objects that are more similar to each other than they are to objects that belong to different groups. The objects may or may not belong to defined groups, but ideally, a clustering algorithm would be able to identify the correct group of examples without access to labels. For example, presented with a sequence of hand-written digits, a clustering algorithm should be able to learn that there are ten types of objects (and also label the images correctly).

In the present study, we wish to cluster 12 by 12 matrices representing the connections between order book events. In principle, a suitable method for learning the underlying structure of the data would involve training a few layers of a convolutional neural network to extract potential features of the weight matrices. Convolutional neural networks achieve state-of-the-art performance on image recognition tasks, but also rely on large datasets. With a sample size of just 401 stocks, and with each observation having a dimension of 144 features, such an approach is infeasible. Instead, I pursue a less elaborate, linear method based on principal component analysis: instead of attempting to cluster the data in its original space, I project the data onto a low-dimensional space determined by principal component analysis and try to groups

in the low-dimensional data.

Figure 4.9 displays the results. The top panel of the figure presents the first and second principal components for reference; the bottom panel displays the proportion of the variance explained by the top twelve principal components as well as a scatterplot of the projection of each connection weight matrix onto the subspace spanned by the first two principal components.

4.4.4 Likelihood & Stability

The features identified by the model thus far are encouraging, but do they explain the data well? Next, I briefly examine the issues of model fit and stability. Log-likelihood measures model fit. More precisely, I compare the log-likelihood of model estimates to a standard (homogeneous) Poisson process with a constant intensity vector, which I refer to as the baseline model. The maximum of the absolute values of the eigenvalues of the connection matrix determines the stability of the model. Models with stability value higher than one are unstable: with positive probability, they generate an infinite number of events in any finite amount of time. At the other extreme, a model with a stability value of zero is a homogeneous Poisson process.

Figure 4.3 compares the fit of the network model versus the baseline model. For each stock, the point-estimator of the network model is the approximate posterior median based on the MCMC sampling algorithm outlined above. Setting the weight matrix to zero and computing the approximate median of the posterior distribution by

direct sampling produces a Bayesian point-estimator of the baseline model. I compute the values in the histogram shown as follows. For each stock, I randomly select 100 subsamples consisting of 180 events each from the same dataset used to construct the point estimates. On each of the subsamples, I compute the ratio of the log-likelihood of the network and baseline models divided by the number of events in the sample and record the average of this value across subsamples. The resulting histogram shows that the network model improves on the baseline model by approximately 4 "bits per event"; the minimum improvement is 1.5 bits per events, and the maximum exceeds 7. These results quantify the extent to which a complex interaction of events determines order book dynamics.

The range of values in Figure 4.3 demonstrates that the network model fits some stocks better than others. Figure 4.4 compares the fit of the model to characteristics of each stocks order book. Specifically, for each stock, I use the same message data used to identify order book events to calculate four attributes of the message data. The characteristics are defined as follows: $Volume$ is the total number of shares traded as identified by execution messages, including executions against hidden orders; $Price$ is the average midprice across all order book updates; $Volatility$ is the standard deviation of $Price$; $Spread$ is the average bid-ask spread across all order book updates. The scatterplots in Figure 4.4 demonstrate that the quality of the network model fit depends on these characteristics. The model appears to work better for stocks with higher trade volumes, lower average prices, lower price volatility, and smaller bid-ask spreads.

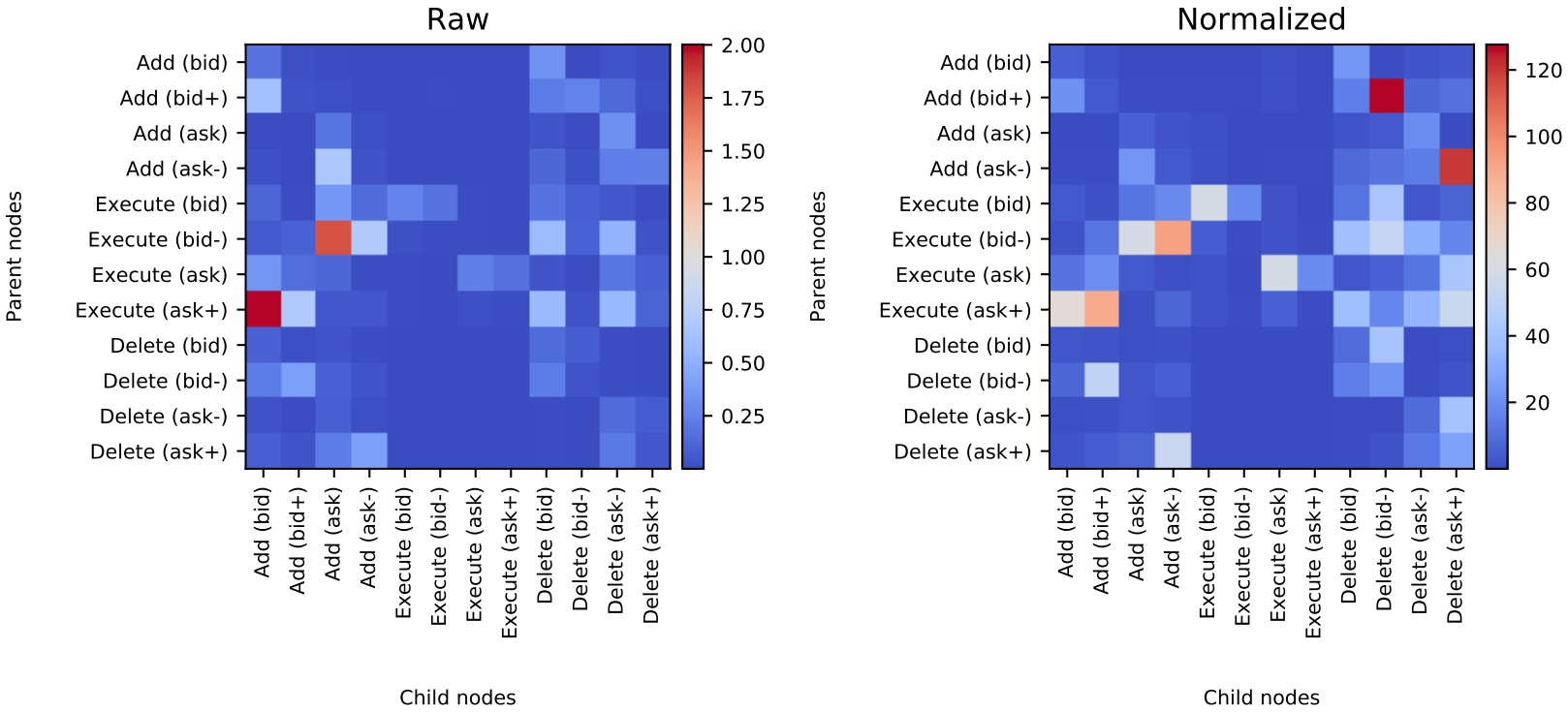
Confirming model stability is essential for two reasons. First, because the network model is only physically meaningful if it is stable. Second, because an unstable model cannot serve as a useful order book simulator. Figure 4.3 presents a histogram of the stability values based on median point estimates. The median value is approximately 0.85, which is well below both the instability threshold (1.0) and estimates reported in studies of related models (e.g., REF). Two of the models have unstable dynamics according to this measure. Figure 4.5 shows scatterplots of stability versus the microstructure characteristics from Figure 4.4. Stability appears to decrease with lower trading volume but is otherwise uncorrelated with the remaining variables. Overall, these results demonstrate that the model learns parameter values leading to highly inhomogeneous, but stable, dynamics.

The results thus far demonstrate that including interactions between events improves model fit relative to a baseline model. They also show that some interaction terms are stronger than others, in both an absolute and relative sense. We can further analyze the importance of interactions terms by answering the following question: How much does the model fit change if I turn off the weaker connections? I answer this question by performing a simple experiment. For a given stock I select a set of threshold values based on the distribution of the estimated matrix weights. Under the minimum threshold, the model includes nearly all of the of the interaction terms, while at the maximum the model reduces to the baseline homogeneous Poisson process. For each threshold, I measure the in-sample and out-of-sample likelihood using subsamples as described above.

Figures 4.6 and 4.7 demonstrate the results for two randomly selected stocks (LLY and SIG). The plots display similar patterns: including weights whose logarithm is less than -4-which amounts to ignoring around half the interaction terms–has little or no effect on the fit of the model. In fact, retaining weights above 0.1 (about a quarter of the interaction terms) achieves the majority of the benefit of the network structure. These observations provide a new perspective on Figure 4.1. Figure 4.8 recreates the plot of the median connection matrix, but this time applies a hard threshold of -2 to the logarithm of the raw weights. The interaction terms that survive are those that were identified earlier, meaning that these are in fact the essential elements of typical order book dynamics.

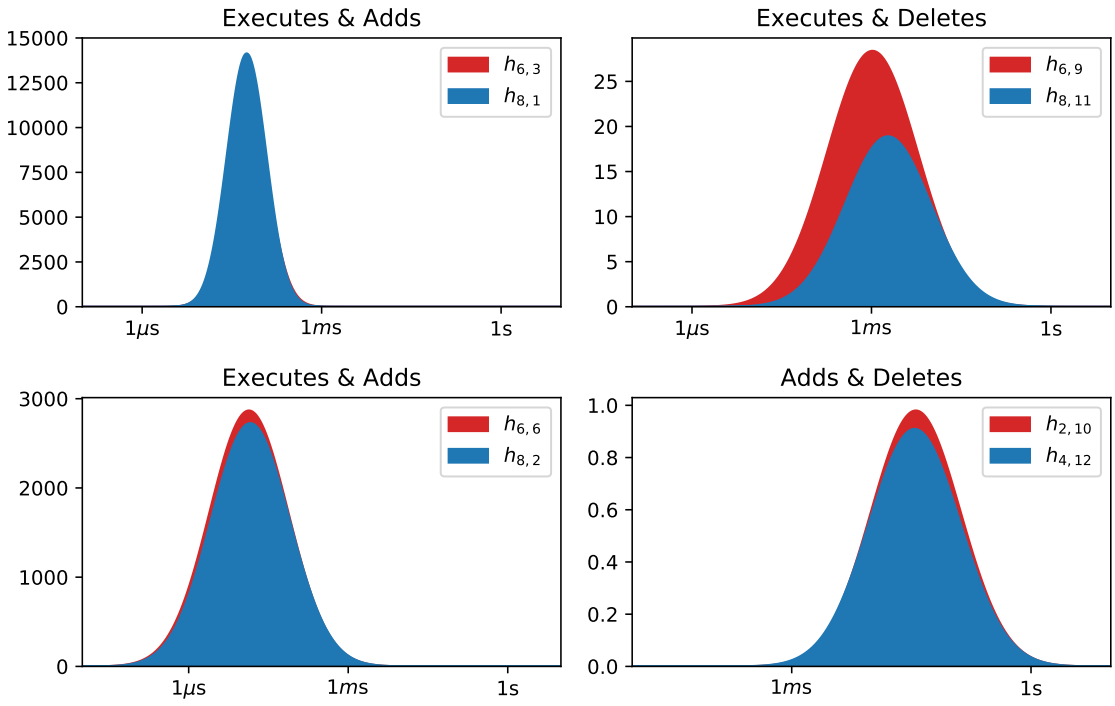
4.5 Conclusion

Figure 4.1: Median Event Connections



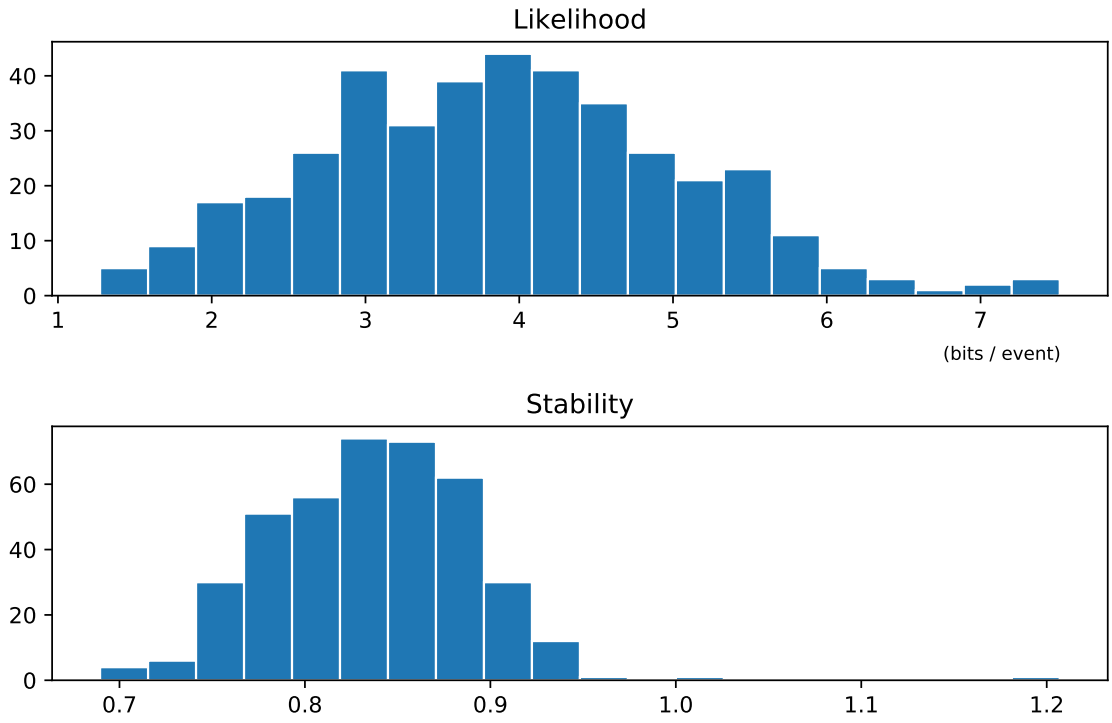
Note: The figure shows the median connection strength between event types across the full sample of stocks. Larger values indicate stronger connections, which are expected to generate a greater number of child events. (Left) The median of connection weight estimates using the raw weight matrix estimates. This figure identifies the most important connections in an absolute sense. (Right) The median connection weight matrix after normalizing parent-child weights by the estimated background rate of child events. This figure identifies the most important connections in a relative sense.

Figure 4.2: Median Impulse Responses



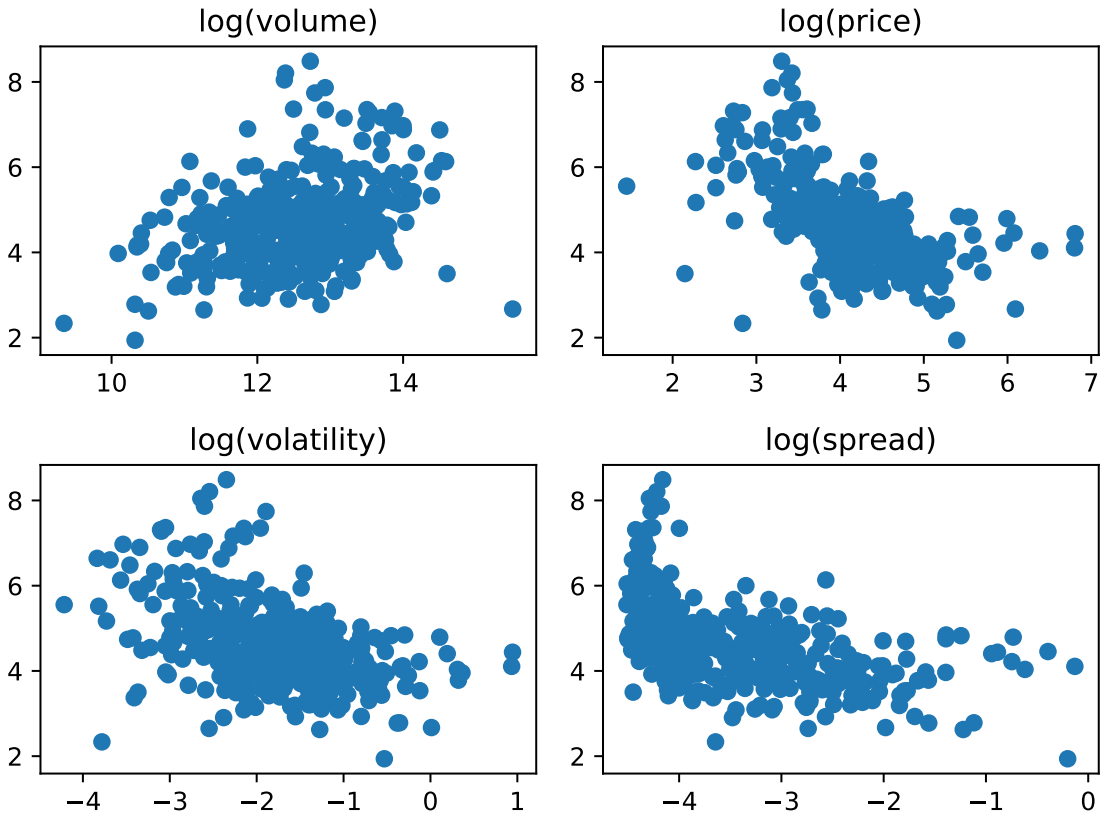
Note: The figure shows the median impulse response for selected parent-child event connection pairs across the full sample. The horizontal axis is logarithmic time to improve interoperability. The figure demonstrates that important connections operate at timescales from one microsecond to one second, that estimates of connections pairs are typically symmetric, and that the model identifies the response to parents as a localized "spike" in the probability of child events.

Figure 4.3: Distribution of Likelihood & Stability



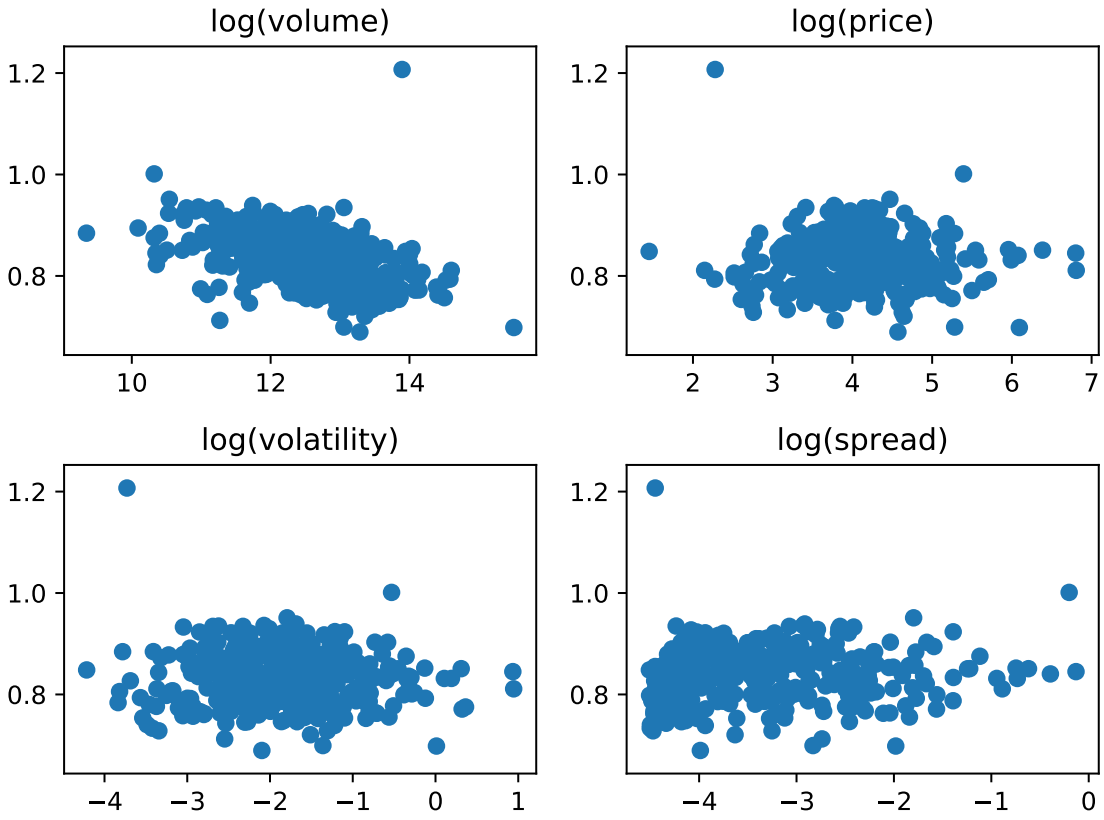
Note: The figure shows histograms of the likelihood and stability of model estimates across the full sample. To compute the likelihood, I select random subsamples from the full sample of daily events on which the model was estimated and average the likelihood across subsamples. The values shown are the differences between the inhomogeneous and homogeneous model divided by the total number of events in each subsample. Stability is computed as the absolute value of the largest eigenvalues of the connection weight matrix. Values below one are stable; values at and above one are unstable. The figure demonstrates that the network model produces a better in-sample fit to the data and that the model estimates are stable in almost all cases.

Figure 4.4: Likelihood vs. Order Book Characteristics



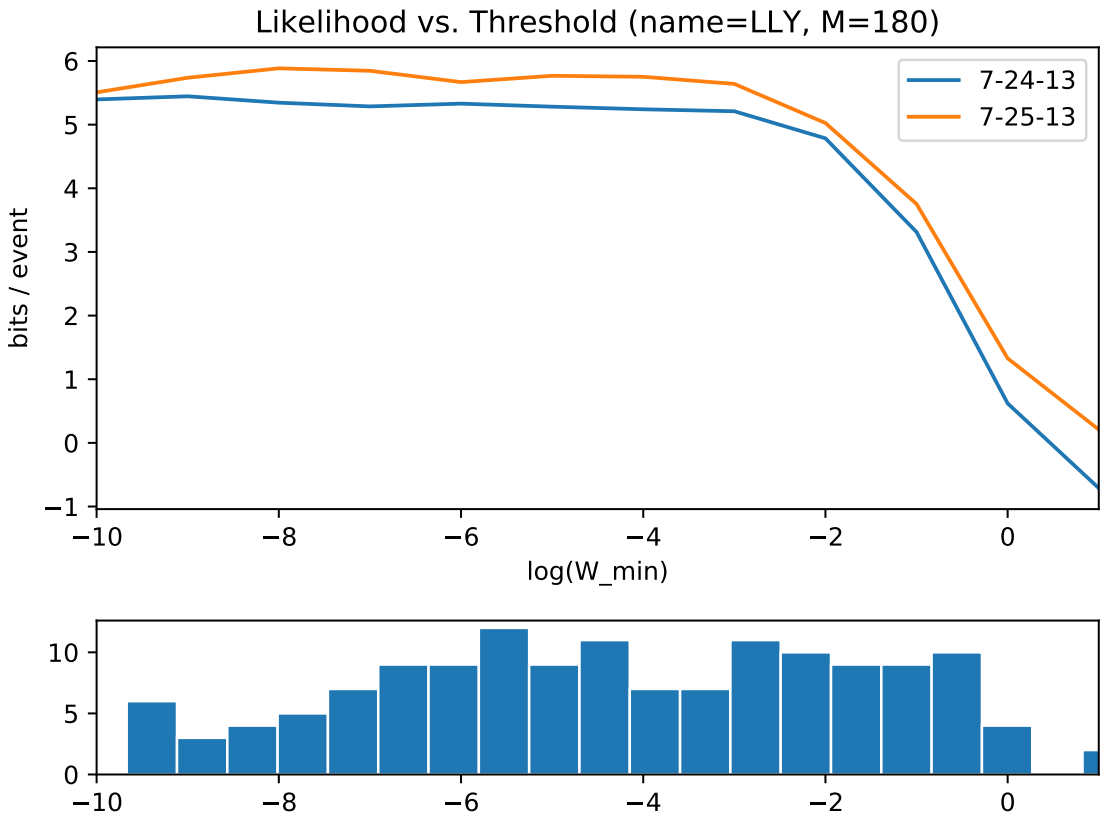
Note: The figure shows scatterplots of the likelihood of estimated models versus properties of the order books calculated from the message data used to identified order book events (i.e., the same day and hours). Volume is the total number of shares traded, including executions against hidden orders; Price is the average midprice across order book updates; Volatility is the standard deviation of Price; Spread is the average bid-ask spread across order book updates. The likelihood is the difference in likelihood between the network model and a standard Poisson model divided by the number of events.

Figure 4.5: Stability vs. Order Book Characteristics



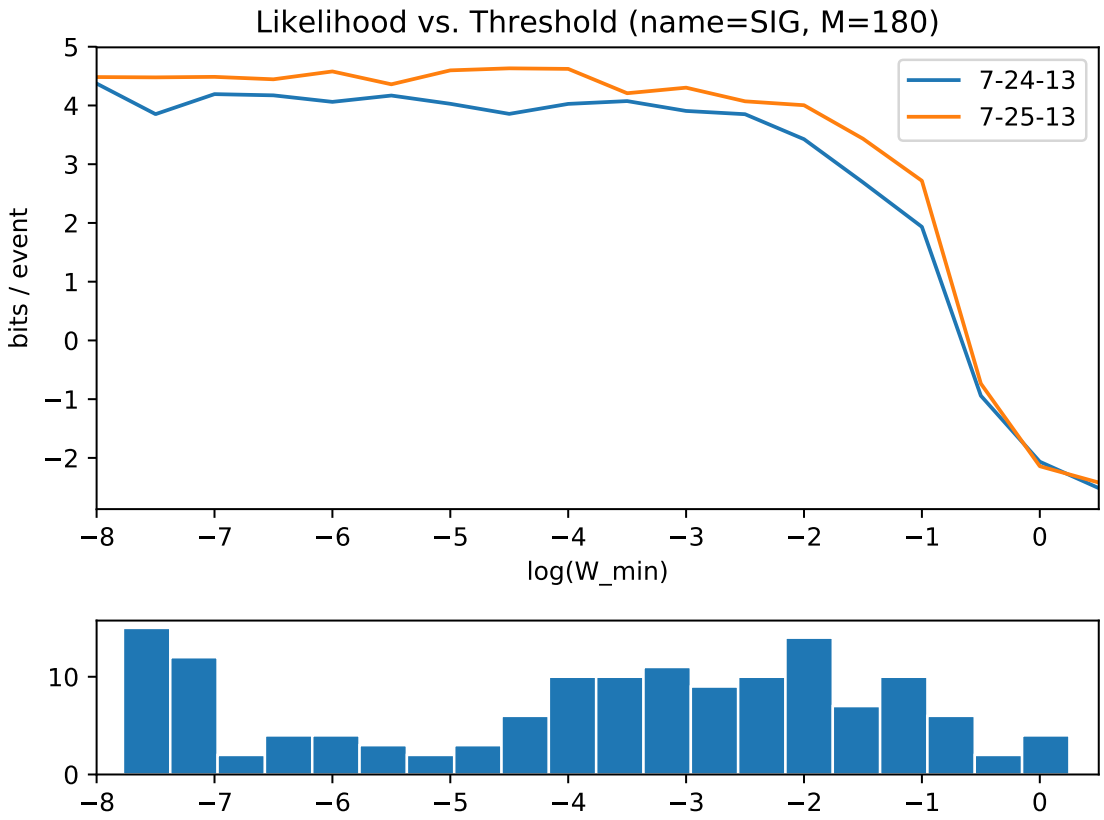
Note: The figure shows scatterplots of the stability of estimated models versus properties of the order books calculated from the message data used to identified order book events (i.e., the same day and hours). Volume is the total number of shares traded, including executions against hidden orders; Price is the average midprice across order book updates; Volatility is the standard deviation of Price; Spread is the average bid-ask spread across order book updates. Stability is the maximum absolute value of the estimated connection weight matrix.

Figure 4.6: Threshold Analysis (LLY)



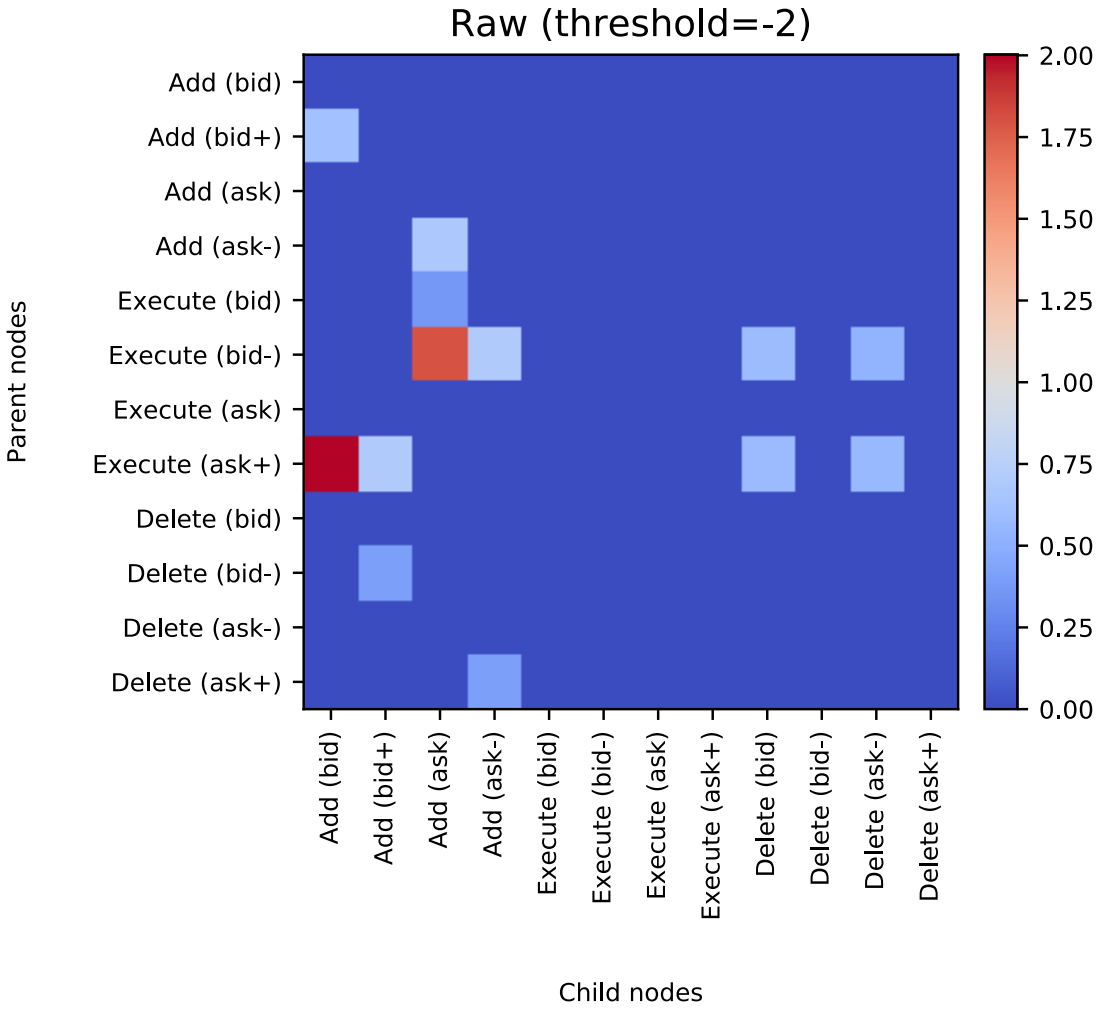
Note: The figure shows the improvement in the likelihood of the network model compared to a standard Poisson model for various thresholds levels. (Top) For each threshold level, the likelihood improvement is computed using the estimated connection weights with all weights below the threshold set to zero. The blue curve shows results using events from the same day the model is estimated on; the orange curve shows results computed on the following day. (Below) A histogram of the connection weights. The figure demonstrates that the ability of the model to fit the data is largely invariant to weights below the median estimated weight and that the model fits out-of-sample data as well as in-sample data.

Figure 4.7: Threshold Analysis (SIG)



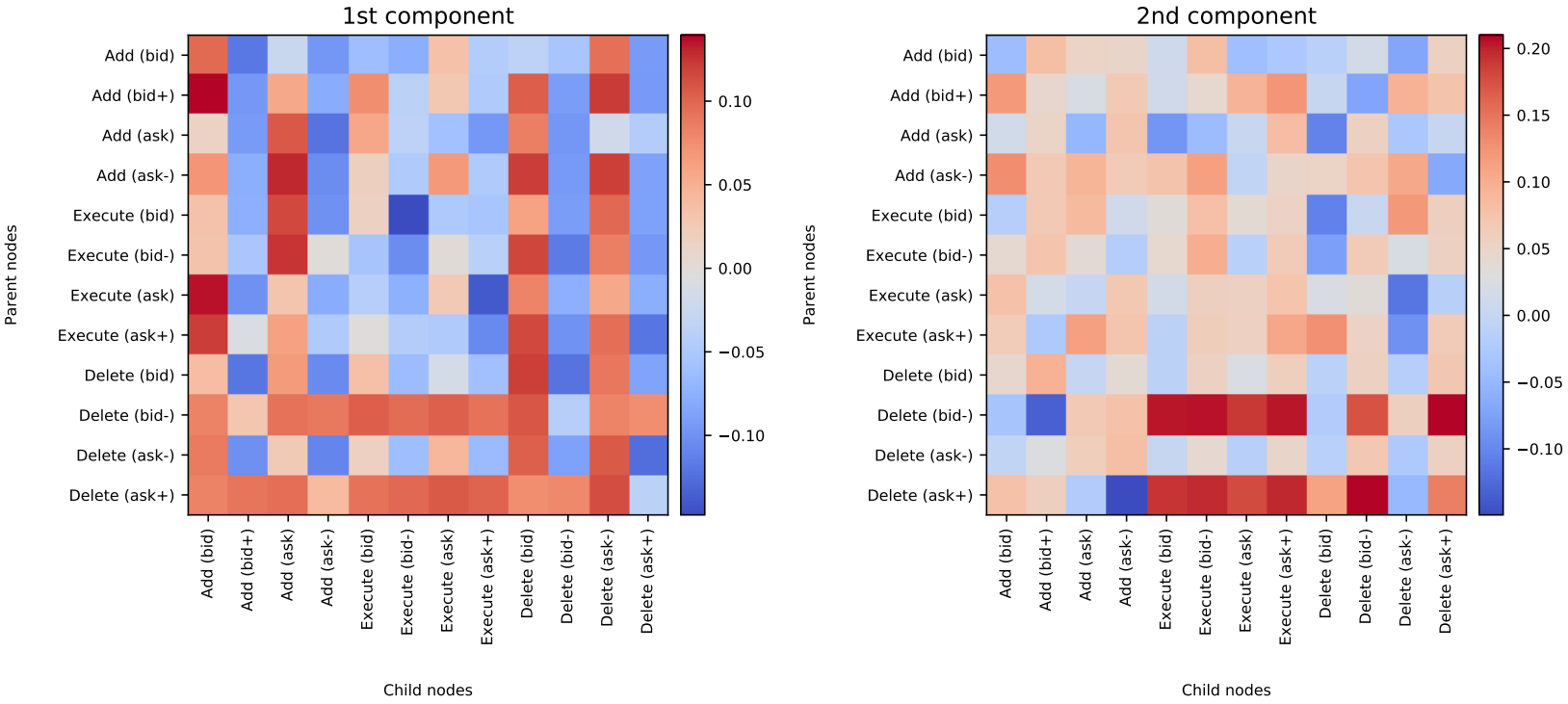
Note: The figure shows the improvement in the likelihood of the network model compared to a standard Poisson model for various thresholds levels. (Top) For each threshold level, the likelihood improvement is computed using the estimated connection weights with all weights below the threshold set to zero. The blue curve shows results using events from the same day the model is estimated on; the orange curve shows results computed on the following day. (Below) A histogram of the connection weights. The figure demonstrates that the ability of the model to fit the data is largely invariant to weights below the median estimated weight and that the model fits out-of-sample data as well as in-sample data.

Figure 4.8: Composite Weights with Threshold



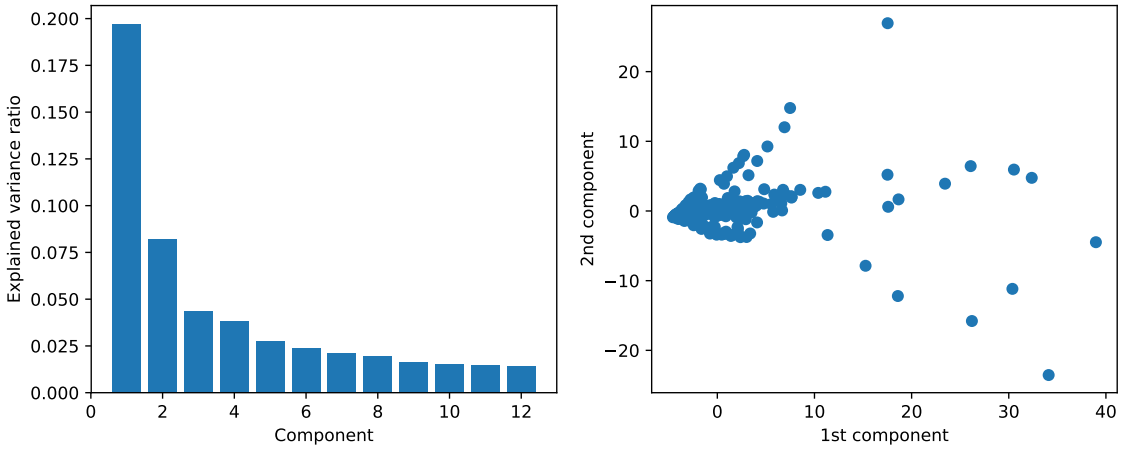
Note:

Figure 4.9: Principal Component Analysis



Note:

Figure 4.10: Principal Component Analysis (continued)



Note:

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.

Yoon Sik Cho, Aram Galstyan, Jeff Brantingham, and George Tita. Latent point process models for spatial-temporal networks. *arXiv:1302.2671*, 2013.

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.

Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(I):83, 1971.

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.

John F. C. Kingman. *Poisson Processes (Oxford Studies in Probability)*. Oxford University Press, January 1993.

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