

ESSAYS IN EMPIRICAL FINANCE WITH LATENT STRUCTURE MODELING

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

Copyright by  
COLIN SWANEY  
2018  
All Rights Reserved

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

To my wife Xiayi and son Aedan.

## ACKNOWLEDGEMENTS

I am indebted to the members of my doctoral committee for their suggestions and insights, and especially grateful to Dr. Artem Durnev, Dr. Jon Garfinkel, and Dr. Qihang Lin for their help and encouragement. This research was supported in part by data provided by Nasdaq, Inc. as well as by computational resources provided by The University of Iowa. Finally, I wish to thank my parents for their love, support, and patience.

## ABSTRACT

This thesis consists of three essays that attempt to provide original empirical analyses of significant problems in finance. The first essay deals with the returns of actively managed mutual funds; the second and third essays attempt to bring further understanding to ultra-high-frequency market microstructure data. The thread that binds these chapters together—if any—is the use of latent structure models. The first chapter attempts to identify unobserved populations of managers, the second looks for hidden structure in order book shapes, and the third searches for connections between order book events.

Evaluating the performance of actively managed equity mutual funds is among the most central topics in the field of finance. In the first chapter, I present a new assessment of the stock picking ability of actively managed funds that accounts for the occurrence of false positives, an issue that complicates traditional assessments. I find that while the data is consistent with a small group of alpha-generating funds, the composition of this population experiences significant annual turnover and is, therefore, difficult to identify in advance. Between 1975 and 2015, the returns to a fund selection strategy based on the classification method fail to generate alpha.

The second chapter begins a study of high-frequency limit order book data. With a view towards exploring the information content of limit orders, as opposed to market orders, I propose a factor model of order book shape. I start by building a unique dataset of Nasdaq limit order books that tracks order activity at ultra

high-frequency. Analyzing over 20,000 stock-days, I find that the limit order book comprises three common factors, which I characterize as level, slope, and curvature. By combining these factors alongside price increments in a vector autoregression, I demonstrate that the factors not only explain limit order book shape but also predict returns over one-minute time intervals. In agreement with the claim that high-frequency traders serve a role in increasing market efficiency, I find a negative correlation between predictability and high-frequency trade activity.

In the third chapter, I explore a continuous-time, event-driven model of limit order book dynamics. It is the first analysis of its kind to examine the microstructure of a broad cross-section of markets, as well as the first to introduce a Bayesian framework for the study of mutually-exciting Poisson processes in order-driven microstructure models. The picture that emerges is that of a strongly self- and mutually-exciting process characterized by intensity “spikes” lasting mere fractions of a second. The largest of these spikes are expected to generate between 0.5 and 2.0 order book events—up to 120 times the number of expected events per second. In the typical order book, market orders demonstrate a significant influence on limit orders and cancellations, but the relationship is non-reciprocating: while limit orders and cancellations exhibit strong interactions with each other, they have no effect on the arrival of market orders. Over 99.5% of the markets examined are stable, and in every market examined, the network model significantly improves in-sample fit relative to a baseline Poisson model. I argue that this improvement is due almost entirely to the most active 20-25% of connections.

## PUBLIC ABSTRACT

This thesis consists of three essays that present original empirical analyses of critical problems in finance. The first essay deals with the returns of actively managed mutual funds; the second and third essays attempt to bring further understanding to the nature of ultra-high-frequency trade data. The thread that binds these chapters together is the use of latent structure models in which the researcher proposes a plausible data model, and then designs statistical methods to learn the details of this structure from data.

Evaluating the performance of actively managed mutual funds is a challenging task. One source of difficulty arises from the sheer number of funds: with thousands of funds, a small group of funds are likely to *appear* to provide value purely by chance. The first essay in this thesis investigates a novel method of accounting for such “false discoveries” in the search for superior returns.

The second and third essays analyze data from public trading on the Nasdaq stock exchange. In the second essay, I demonstrate that fully electronic limit order books possess characteristic “shapes” that predict future price movements. The results of the study reinforce a crucial feature of modern exchanges, which is that limit orders reveal information about traders information. The third and final essay analyzes a new statistical model that attempts to capture essential features of order book activity. I demonstrate that this activity is characterized by a small number intense interactions taking place over brief windows of time.



# TABLE OF CONTENTS

LIST OF TABLES . . . . .	ix
LIST OF FIGURES . . . . .	x
CHAPTER	
1 EVALUATING FUND MANAGER SKILL: A MIXTURE MODEL APPROACH . . . . .	1
1.1 Introduction . . . . .	1
1.2 Evaluating Skill . . . . .	4
1.2.1 Framework for Evaluating Skill . . . . .	4
1.2.2 Estimation Procedure . . . . .	6
1.2.2.1 Advantages of Expectation-Maximization . . . . .	8
1.3 Data . . . . .	9
1.4 Results . . . . .	10
1.4.1 Long Run Skill . . . . .	12
1.4.2 Short Run Skill . . . . .	12
1.4.3 The Evolution of Skill . . . . .	13
1.4.4 The Persistence of Skill . . . . .	15
1.4.5 Estimate Validation . . . . .	22
1.5 Conclusion . . . . .	25
2 PRICE FORMATION AND ORDER BOOK SHAPES . . . . .	28
2.1 Introduction . . . . .	28
2.1.1 Summary . . . . .	30
2.1.2 Contribution . . . . .	32
2.2 Data . . . . .	33
2.3 Order Book Shape . . . . .	37
2.4 Order Book Dynamics . . . . .	44
2.5 High-Frequency Trading . . . . .	49
2.6 Conclusion . . . . .	55
3 ORDER BOOK EVENTS ON A POISSON NETWORK . . . . .	59
3.1 Introduction . . . . .	59
3.1.1 Summary . . . . .	62
3.1.2 Contribution . . . . .	64
3.2 Model . . . . .	66

3.2.1	Mutually-Exciting Poisson Processes . . . . .	67
3.2.2	Order Book Events . . . . .	70
3.2.3	Parameterization . . . . .	72
3.2.4	Inference . . . . .	75
3.3	Data . . . . .	79
3.4	Results . . . . .	81
3.4.1	Connections . . . . .	83
3.4.2	Impulse Response . . . . .	88
3.4.3	Further Analysis . . . . .	91
3.4.3.1	Likelihood . . . . .	93
3.4.3.2	Stability . . . . .	95
3.4.3.3	Model Selection . . . . .	96
3.5	Conclusion . . . . .	99
REFERENCES . . . . .		103

## LIST OF TABLES

Table

1.1	Summary of fund return statistics . . . . .	11
1.2	Mixture model parameter estimates . . . . .	14
1.3	Fund classification proportions . . . . .	17
1.4	Fund classification turnover . . . . .	19
1.5	Fund performance by classification . . . . .	21
1.6	Monte Carlo simulation . . . . .	27
2.1	Vector autoregression estimates . . . . .	48
2.2	The effect of high-frequency trade on predictive regressions . . . . .	56
2.3	Variable descriptions . . . . .	58
3.1	Order book event definitions . . . . .	73
3.2	Summary statistics for ITCH message data . . . . .	80

## LIST OF FIGURES

Figure

1.1	Example of a mixture model distribution . . . . .	5
1.2	Time series of the mixture model parameter estimates . . . . .	16
1.3	Time series of portfolio performance by threshold . . . . .	22
2.1	Sample stock characteristics . . . . .	36
2.2	Examples of average limit order book shapes . . . . .	39
2.3	Principal component analysis of order books . . . . .	42
3.1	Example of a mutually-exciting Poisson process. . . . .	71
3.2	Examples of logistic-normal impulse responses . . . . .	74
3.3	Patterns in order book message activity . . . . .	82
3.4	Median connection weight point estimates . . . . .	86
3.5	Median impulse response point estimates . . . . .	90
3.6	Example of a daily fitted intensity: SIG . . . . .	92
3.7	Distributions of likelihood & stability . . . . .	94
3.8	Likelihood versus order book characteristics . . . . .	97
3.9	Stability versus order book characteristics . . . . .	98
3.10	Likelihood versus connection weight threshold: LLY . . . . .	100
3.11	Likelihood versus connection weight threshold: SIG . . . . .	101
3.12	Median connection weight point estimates with threshold . . . . .	102

# CHAPTER 1

## EVALUATING FUND MANAGER SKILL: A MIXTURE MODEL APPROACH

### 1.1 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. This chapter provides an alternative perspective on the results found in Barras et al. (2010). The 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 relate to an *unsupervised learning* task, as 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 Monte Carlo simulation. 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.

## 1.2 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: I label these the *poor*, *average*, and *good* groups. 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.

### 1.2.1 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 and that these differences 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



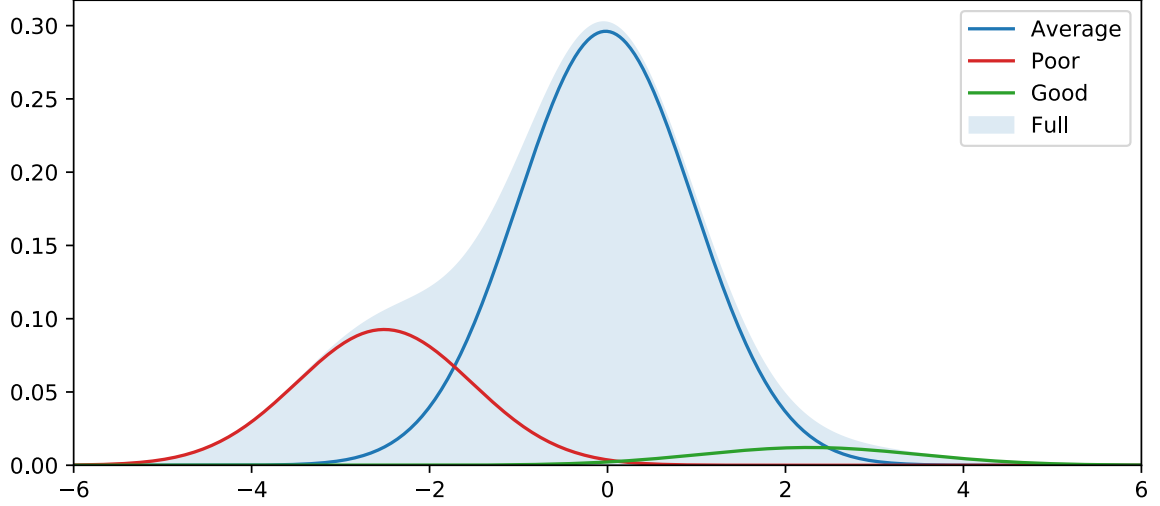


Figure 1.1. Example of a mixture model distribution. The shaded region represents the observed distribution; solid curves represent the underlying, hidden sub-populations. For the distribution shown, the proportions of the sub-populations come from Barras et al. (2010): 23% (red), 73.5% (blue), and 3.5% (green).

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 and fails at the latter if the number of funds in a particular subpopulation is small, or if manager performance depends on unobserved, time-varying parameters. From an investment standpoint, the first task may be relevant even if the later proves difficult because optimal asset allocation (from a Bayesian perspective) depends on an investor's prior beliefs about investment opportunities.

The specific model I have in mind is the mixture of normals model depicted in Figure 1.1. With unconditional probability  $\pi_j$ , observation  $\alpha_i$  is drawn from population  $j$ . Observations of alpha from population  $j$  have a normal distribution with mean  $\mu_j$  and standard deviation  $\sigma_j$ . 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  $\mu_j$ ,  $\sigma_j$ , and  $\pi_j$  ( $j = 1, 2, 3$ ). Referring to Figure 1.1, she observes the full distribution but believes that the underlying sub-populations exist and attempts to learn the parameters describing them.

### 1.2.2 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 a fMRI 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. I outline the basic algorithm below.

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  $\Theta$ , 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  $\Theta$ . I repeat steps two and three until the estimate of  $\Theta$  converges (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,  $\Theta$  consists of the means, standard deviations, and weights assigned to each skill population:

$$\Theta = \begin{pmatrix} \mu_1 & \mu_2 & \mu_3 \\ \sigma_1 & \sigma_2 & \sigma_3 \\ \rho_1 & \rho_2 & \rho_3 \end{pmatrix} \quad (1.1)$$

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 (Robert and Casella (2010)). Moreover, the estimation procedure generates point estimates of the conditional probabilities that observation  $i$  obtains from population  $j$ ,

$$p_j^i = f_{y_i|x_i, \Theta}(y_i = j) \quad (1.2)$$

The estimates yield a simple classification scheme, as each observation naturally identifies 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  $\Theta_{mle}^*$  minimizes the mean squared error amongst all consistent estimators.

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.

### 1.2.2.1 Advantages of Expectation-Maximization

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. 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  $\alpha$ . 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*, *Poor*, or *Average* performers.

### 1.3 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 2015. 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.

The returns data was used to construct a collection of regression intercept estimates (i.e., alphas) and corresponding  $t$ -statistics. I generate these values using

the four-factor model studied by Carhart (1997):

$$r_{i,t} = \alpha_i + b_i \cdot r_{m,t} + s_i \cdot r_{smb,t} + h_i \cdot r_{hml,t} + m_i \cdot r_{mom,t} + \epsilon_{i,t}, \quad (1.3)$$

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 1.1 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)).

## 1.4 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 ability. Finally, I validate my results through a simulation exercise.

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

Table 1.1. Summary of fund return statistics. 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 quantiles for the distribution of fund  $t$ -statistics using the full sample period.

### 1.4.1 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 the previous section. Table 1.2 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 means of the worst performing class and the best performing class (-2.33 versus 1.08, respectively). While the estimated difference in means refers to  $t$ -statistics and not returns, it is clear that there is a substantial cost to investing in these poorly performing funds.

### 1.4.2 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 1.2 shows the maximum likelihood estimates for the short-run sample. 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 are less likely over the long-run.

### 1.4.3 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  $\mu_t$ ,  $\sigma_t$ , and  $\rho_t$  running from 1990 to 2015.

Figure 1.2 shows the results. The time series of estimated means fails to 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 differ from the time series results in Barras et al. (2010). They find a steady downward trend in the size of the skilled population, as well as a corresponding increase in the group of unskilled funds. While a reduction in the size

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

Table 1.2. Mixture model parameter estimates. The table shows the parameter estimates for the distribution of mutual fund  $t$ -statistics. *Long-Run* results were obtained using each funds full history of monthly returns. *Short-Run* results were obtained by first dividing each return history into five-year sub-samples, then estimating (1.3) on each sub-sample. Values in parentheses are bootstrapped standard errors.

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.

#### 1.4.4 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  $\tau = 0.5, 0.6, \dots, 0.9$ . The classification rule is simple: if the probability that the fund belongs to class  $j$  exceeds the threshold, then it is assigned to that class (the threshold choices imply that the fund is assigned to at most one class). If no good funds are found at a particular threshold, then the returns over the holding period for that portfolio are substituted by the monthly market returns ( $\alpha = 0$ ). Otherwise, the portfolio is held for one year, after

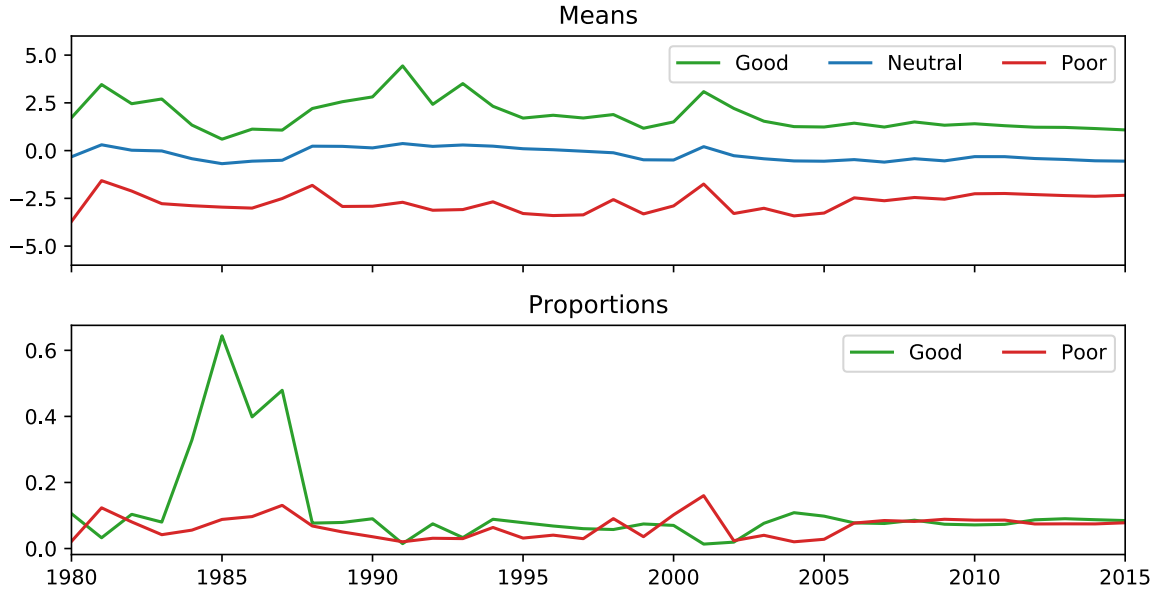


Figure 1.2. Time series of the mixture model parameter estimates. (Top) For year from 1980 to 2015 I plot the time series of the estimated populations means of the mixture model for  $t$ -statistics using fund returns up to that year. (Bottom) The estimated populations weights of the mixture model for  $t$ -statistics using fund returns up to that year.

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 1.3 describes the composition of the resulting portfolios using the ratio of funds assigned to each portfolio. The proportion of funds in each portfolio is ordinarily smaller than the estimated portion 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 ( $\tau > 0.9$ ).

Table 1.4 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

Population	$\tau$	$\bar{\pi}_c$	Proportion				
			$= 0\%$	$0 - 6\%$	$6 - 12\%$	$12 - 24\%$	$> 24\%$
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

Table 1.3. Fund classification proportions. 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 *Poor*, *Average*, 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 2015. 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 *Average* if it is not assigned to any of the *Poor* 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.

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 1.5 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 strictly increasing in  $\tau$  for the good portfolios, nor are they decreasing for the poor portfolios. One reason for this is that 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, the classification procedure provides a benefit to an investor choosing amongst actively managed funds, but only through its ability to *avoid* poor performing funds.

Figure 1.3 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 1.2. Whereas the estimated group means fluctuate randomly, the portfolio alphas steadily decrease during this period. As neither the size nor the ability of the best performing

Population	$\tau$	Years After Classification				
		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

Table 1.4. Fund classification turnover. 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 *Poor*, *Average*, 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 *Average* if it is not assigned to any of the *Poor* 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  $\sigma$  are annualized.

population declined over this part of the sample, this finding suggests that the good performers became less *consistent*, as reflected by the high turnover documented in Table 1.4.

Care is required when interpreting the results shown in Figure 1.3. What the figure shows is that before 1990 the average performance of the best portfolio was positive and statistically significant. 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 (statistically) 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 fluctuations 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 1.4 we know that the composition of portfolios changes frequently and dramatically. Instead of measuring performance by alpha as in Table 1.5, 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 from a four-factor model,  $\mathbf{B}_t$ , and compute the monthly alphas for each of the next twelve months as  $\alpha_{t+i} = r_{t+i} - \mathbf{B}_t \mathbf{f}_{t+i}$  ( $i = 1, \dots, 12$ ). This procedure results in the time series of monthly alphas displayed in Table 1.5. The average annualized



Population	$\tau$	$\hat{\alpha}$	$p$	$\hat{b}_{mkt}$	$\hat{b}_{smb}$	$\hat{b}_{hml}$	$\hat{b}_{umd}$	$\bar{r}_p$	$\sigma_p$
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

Table 1.5. Fund performance by classification. 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 *Poor*, *Average*, 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 *Average* if it is not assigned to any of the *Poor* 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  $s = 1, \dots, 5$ .

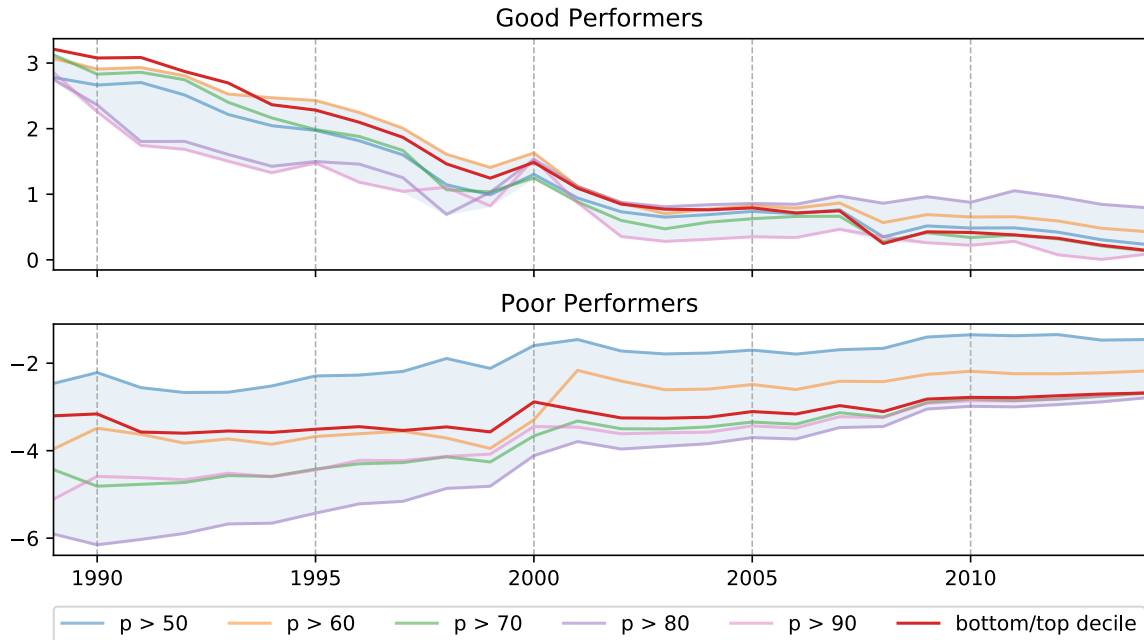


Figure 1.3. Time series of portfolio performance by threshold. At the beginning of each year from 1989 to 2014 I estimate a four-factor alpha for each good and poor performing fund using the entire history of monthly portfolio returns up to that date. The top (bottom) figure compares the performance of good funds with an alternative investment strategy that invests in the top (bottom) decile of funds determined by each funds four-factor t-statistic over the preceding 60-month period.

monthly alphas of portfolios of poor and average performers are substantially lower than those of portfolios of good performers, and all portfolios of good performers have positive average alphas. The results agree with Table 1.5: despite high fund turnover, the average performance of the classified portfolios agrees with their type and the estimates in Table 1.2.

#### 1.4.5 Estimate Validation

Expectation-maximization produces numerical approximations to maximum likelihood estimates. Theoretically, maximum likelihood estimates have desirable

Population	$\tau$	$\bar{\alpha}_t$	$\sigma(\alpha_t)$	$q_{0.05}$	$q_{0.10}$	$q_{0.50}$	$q_{0.90}$	$q_{0.95}$
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

Table 1.5 (continued). 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 *Poor*, *Average*, 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 *Average* if it is not assigned to any of the *Poor* or *Good* populations with the probabilities shown. Monthly portfolio alphas ( $\alpha_t$ ) 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  $\alpha_t$ . All values are annualized.

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 small, or if the densities of the individual distributions have a significant 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-normals 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 1.6 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 was not found to be statistically different from zero).

These results present a dilemma. If their mixture of normals model is correct, then the estimated weights in the present paper should be similar to those found in Barras et al. (2010). Table 1.2, Panel B shows that—using the same sample period as 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.

## 1.5 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. The results reveal that over an extended sample period the distribution of ability is consistent with the existence of populations of both poor and good performing funds, as well as a majority population of managers that underperforms low cost, passive investment alternatives. I find mixed results concerning the evolution of the distribution of skill. Formal estimates indicate that this distribution has remained stable over time, but the performance of an investment strategy based on these estimates demonstrate a decline in the ability of a portfolio designed to capture the superior performance to outperform passive investment over the latter half of the sample period. I validate these results using a Monte Carlo simulation exercise.

What are we to conclude of the active management industry? On the one hand, it appears that there are 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 challenging because the size of the talented pool is small, its expected excess return is not much larger than that of the average

fund, and the composition of the group is inconstant. For an investor in this asset class, the easier task is avoiding underperforming funds.

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

Table 1.6. Monte Carlo simulation. 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 2 PRICE FORMATION AND ORDER BOOK SHAPES

### 2.1 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—that is, 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,  $m_t = \frac{1}{2} (p_{t,1}^{(bid)} + p_{t,1}^{(ask)})$ . 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  $(w_1, \dots, w_K)$ , 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  $t = 1, \dots, T$ . 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

$$r_t = \sum_k \beta_k \left( \sum_i v_{t,i} w_{k,i} \right) = \sum_i v_{t,i} \sum_k \beta_k w_{k,i} = \sum_i v_{t,i} \tilde{\beta}_i, \quad (2.1)$$

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 is a de-noised version 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.

### 2.1.1 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 describing updates to limit order books. 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 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.

### 2.1.2 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 are commonalities 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 critical basis function of the limit order book: its *level*.

In contrast to these studies, I show a small number of statistical factors describes the shapes of limit order books and that each of these factors contains distinct information about future price movements (they are orthogonal by construction). Beltran-Lopez et al. (2009) takes a similar approach, 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.

## 2.2 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 extensive, 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 2.1, 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.

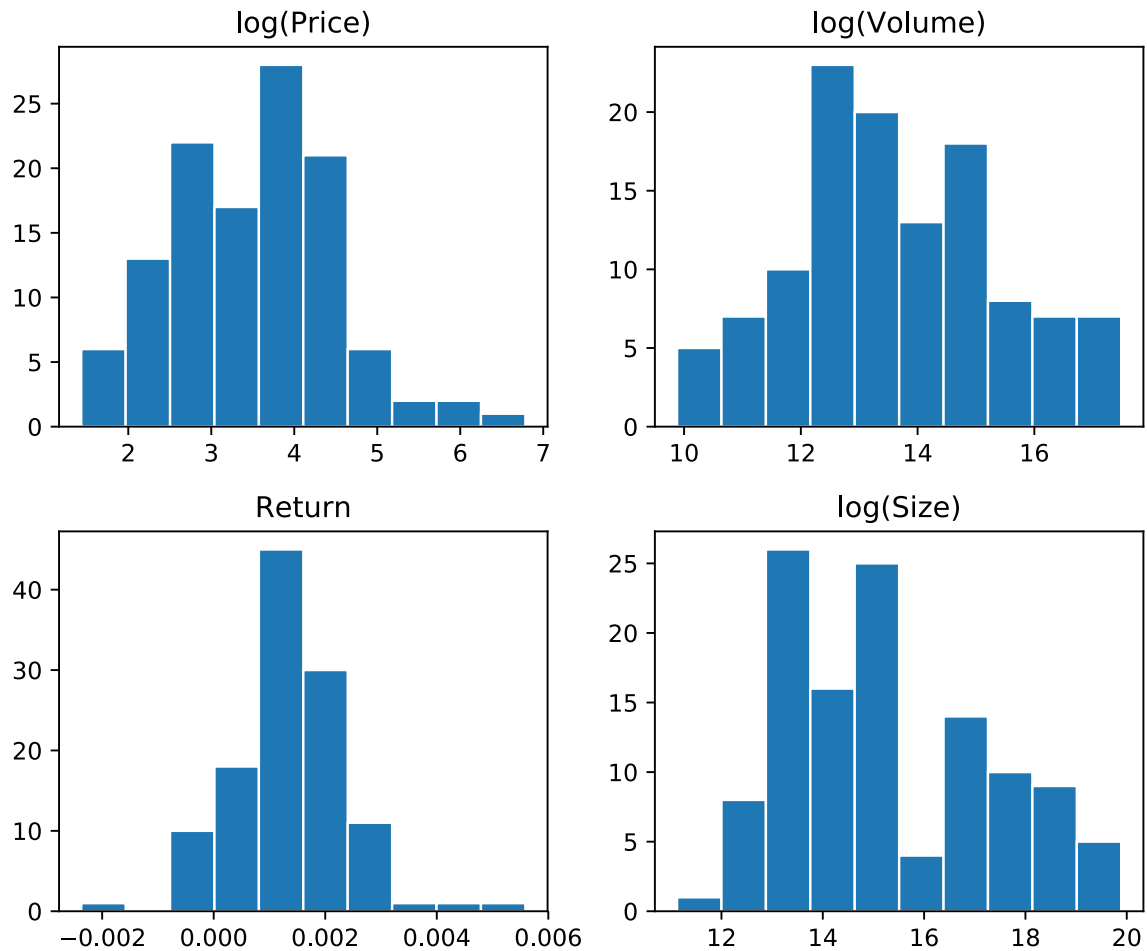


Figure 2.1. Sample stock characteristics. These figures demonstrate the distribution of daily average stock characteristics for my sample.  $\log(\text{Price})$  is the logarithm of the daily closing price;  $\log(\text{Volume})$  is the logarithm of the daily number of shares sold;  $\text{Return}$  is the daily holding period return in basis points;  $\log(\text{Size})$  is the logarithm of the daily market capitalization in thousands. Averages for each stock are calculated from all days in the period Jan. 1, 2013 - Dec. 31, 2013 for which complete limit order book data is available.



A 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 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.

### 2.3 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 2.2 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 average 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, 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-

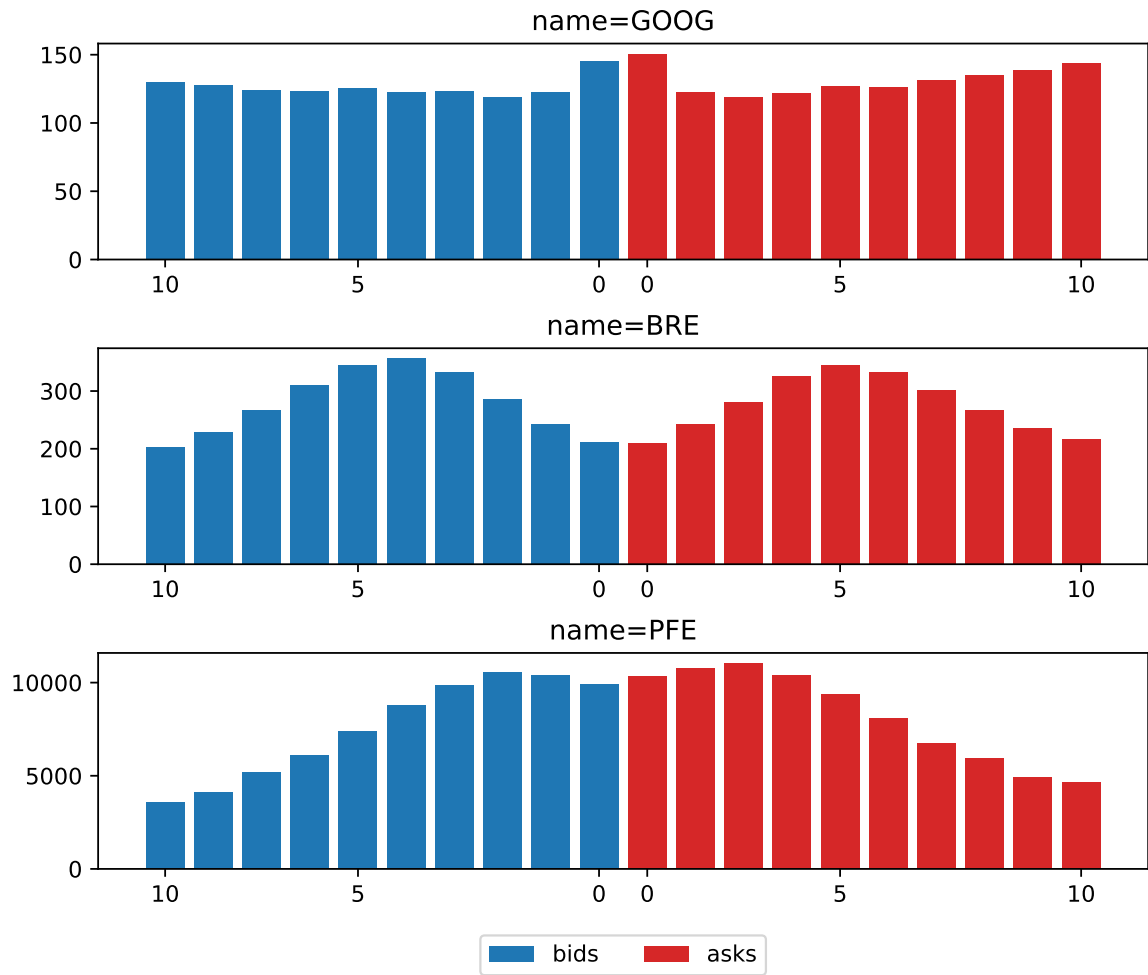


Figure 2.2. Examples of average limit order book shapes. The figure shows the average shape of randomly selected large (top row), medium (middle row), and small (bottom row) firms. Average shapes are calculated from all order book updates in the ITCH database during 2013.

principal component analysis to order book snapshots. Principal component analysis is a method of dimensionality reduction; instead of describing order books using ten price levels, we wish to explain them 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 ( $T \gg N$ ), 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 way to perform principal component analysis is to compute the eigenvalue decomposition of the sample covariance matrix  $(X - \bar{X})^T(X - \bar{X}) = W\Lambda W^T$ , and then define  $P = (X - \bar{X})W$ . In this case the eigenvalues (the diagonal of  $\Lambda$ ) 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  $U_k, \Sigma_k, V_k$  are matrices consisting of the first  $k$  columns of  $U, \Sigma$ , 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  $390 \times 10$  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 2.3 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%.

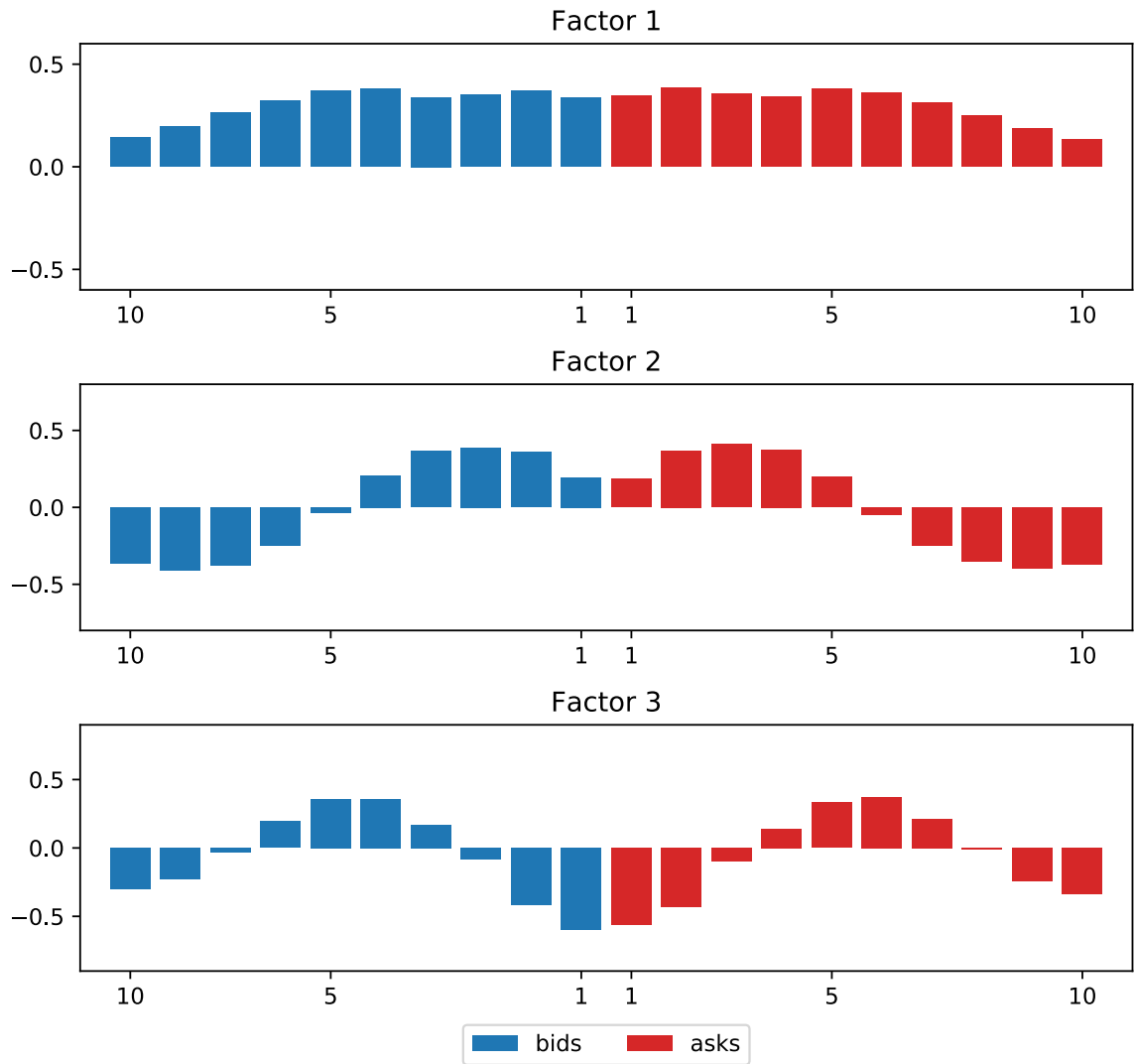


Figure 2.3. Principal component analysis of order books. The figures show the results of the principal component analysis of daily limit order book volume data. For each day in the sample I perform PCA on the daily order book history. I then calculate the average of the stock-day eigenvectors (*top*) 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.

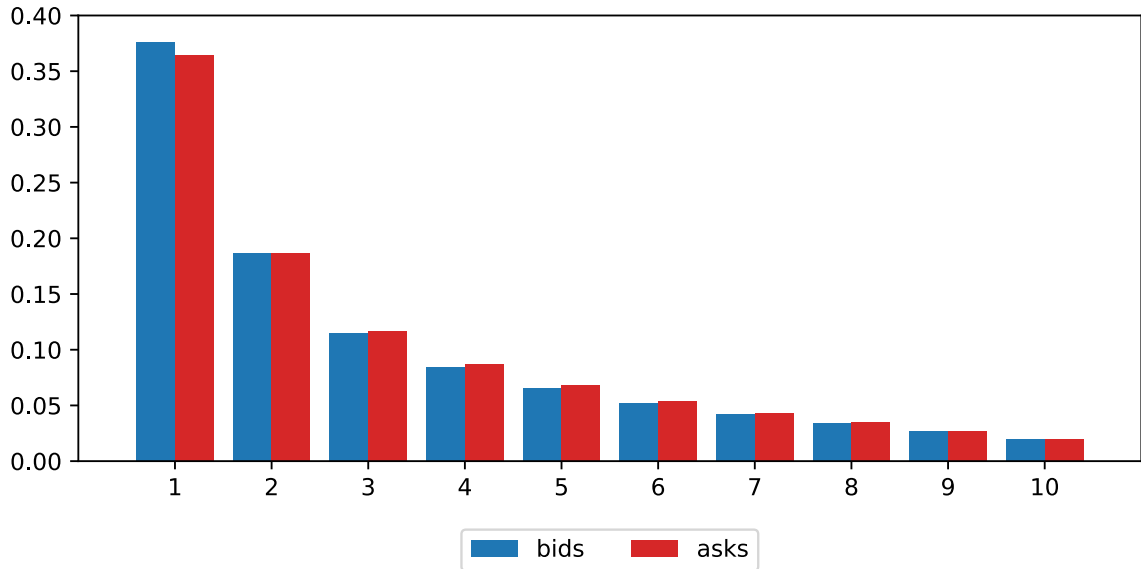


Figure 2.3 (continued). The figures show the results of the principal component analysis of daily limit order book volume data. For each day in the sample I perform PCA on the daily order book history. I then calculate the average of the stock-day eigenvectors (*top*) and the average of the stock-day eigenvalues (*bottom*). Averages are calculated from all order book updates in the ITCH database during 2013 for the stocks in my sample.

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 2.3. In the next section, I analyze the information content these proposed factors.

## 2.4 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. In particular, I estimate a vector autoregressive model of returns and first-differences in the proposed order book factors:

$$X_t = c + \sum_{k=1}^K A_k X_{t-k} + \epsilon_t, \quad (2.2)$$

where

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

I replace the factors by their first differences because they 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 essential features of that system.

The most interesting equation in (2.2) is the return equation

$$\begin{aligned}
 Ret_t = c + \sum_{k=1}^K & \beta_{1,k} Ret_{t-k} + \beta_{2,k} \Delta Lvl_{t-k}^{Bid} + \beta_{3,k} \Delta Lvl_{t-k}^{Ask} \\
 & + \beta_{4,k} \Delta Slp_{t-k}^{Bid} + \beta_{5,k} \Delta Slp_{t-k}^{Ask} + \beta_{6,k} \Delta Crv_{t-k}^{Bid} + \beta_{7,k} \Delta Crv_{t-k}^{Ask}, \quad (2.3)
 \end{aligned}$$

the coefficients of which reflect the predictability of returns based on changes in the shape of the order book. It is useful to consider what signs we might expect for the coefficient estimates in (2.3). 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 part 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 different signs, it 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 coefficient.

Equation (2.2) 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 2.1, 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 2.1 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 the 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 significance in explaining the shape of the limit order book.

What about the dynamics of the order book factors themselves? Panels B-D of Table 2.1 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

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

Table 2.1. Vector autoregression estimates. 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.

equal to the size 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 in the bid (ask) side. For the slope equations, the relationship reverses.

## 2.5 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

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

Table 2.1 (continued). This table shows the average coefficient estimates of the vector autoregressive model of the limit order book:  $X_t = \sum_{k=1}^5 A_k X_{t-k}$ . The components of  $X_t$  are one-minute mid-price returns and the changes to the level, slope, and curvature factors for each side of the order book. The table presents the stock-day average of coefficient estimate and  $t$ -values based on double-clustered standard errors. Coefficient estimates for the shape factors are scaled by  $10^7$ . Coefficient estimates of returns are scaled by  $10^4$ . The estimation period is Jan-2013 to Dec-2013, and the sample excludes stock-days with missing observations. \*\*\*, \*\*, \* indicate significance at the 1%, 5%, and 10% levels, respectively.

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

Table 2.1 (continued). 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.

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

Table 2.1 (continued). 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.



high-frequency trading has an overall beneficial effect of market quality<sup>1</sup>.

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 strategy counts the occurrence of simple patterns in message data that 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 are less than 100 milliseconds apart, (2) they 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

---

<sup>1</sup>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.

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 markets 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 2.2 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 explaining future returns, the overall effect is a reduction in the immediate price impact of increases to these factors.

## 2.6 Conclusion

In this paper, I have attempted to provide a simple representation of electronic limit order book dynamics. My central 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.

	$Ret_{t-1}$	$\Delta Level_{t-1}^B$	$\Delta Level_{t-1}^A$	$\Delta Slope_{t-1}^B$	$\Delta Slope_{t-1}^A$	$\Delta Curve_{t-1}^B$	$\Delta Curve_{t-1}^A$
<i>Intercept</i>	-0.22*** (-5.91)	14.07** (2.05)	-20.75*** (-5.54)	6.86*** (3.57)	-9.82*** (-5.23)	-4.40*** (-3.38)	2.41*** (2.86)
<i>HFT<sub>top</sub></i>	0.03*** (3.606)	-3.08*** (-3.64)	2.07*** (3.20)	0.12 (0.38)	-0.30 (-1.21)	0.00 (0.02)	0.03 (0.20)
<i>log(Price)</i>	0.02** (2.42)	0.22 (0.18)	-0.20 (-0.35)	0.18 (0.71)	-0.11 (-0.38)	0.28 (1.02)	-0.16 (-1.22)
<i>log(Volume)</i>	0.03*** (5.02)	0.19 (0.13)	0.28 (0.52)	0.01 (0.06)	0.38 (1.36)	0.41 (1.26)	-0.14 (-1.38)
<i>Ret</i>	-0.03 (-0.35)	4.86 (0.34)	-6.06 (-0.51)	4.96 (0.53)	2.80 (0.58)	2.65 (0.52)	-1.71 (-0.57)
<i>log(Size)</i>	-0.02*** (-3.49)	-0.78 (-0.69)	0.82* (1.80)	-0.44** (-2.29)	0.26 (1.12)	-0.18 (-0.67)	0.03 (0.26)

Table 2.2. The effect of high-frequency trade on predictive regressions. 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 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.

	$Ret_{t-1}$	$\Delta Level_{t-1}^B$	$\Delta Level_{t-1}^A$	$\Delta Slope_{t-1}^B$	$\Delta Slope_{t-1}^A$	$\Delta Curve_{t-1}^B$	$\Delta Curve_{t-1}^A$
<i>Intercept</i>	-0.36*** (-12.69)	26.98*** (5.57)	-32.91*** (-10.20)	3.59* (1.80)	-7.14*** (-5.08)	-3.70*** (-4.26)	1.49** (2.18)
<i>log(HFT)</i>	0.01*** (4.47)	-1.21** (-2.57)	0.83*** (3.84)	0.37*** (2.93)	-0.28** (-2.51)	-0.14 (-1.42)	0.09* (1.66)
<i>log(Price)</i>	0.02** (2.59)	0.20 (0.16)	0.49 (0.95)	0.37 (1.34)	-0.25 (-0.97)	0.29 (0.93)	-0.10 (-0.78)
<i>log(Volume)</i>	0.04*** (7.24)	-0.57 (-0.44)	1.62*** (3.33)	0.38 (1.14)	0.11 (0.57)	0.36 (1.13)	-0.02 (-0.24)
<i>Ret</i>	-0.02 (-0.29)	5.84 (0.45)	-5.58 (-0.47)	3.66 (0.38)	3.35 (0.66)	2.59 (0.54)	-2.05 (-0.73)
<i>log(Size)</i>	-0.02*** (-3.90)	-0.65 (-0.58)	0.05 (0.11)	-0.71*** (-2.75)	0.44** (2.24)	-0.13 (-0.47)	-0.06 (-0.51)

Table 2.2 (continued). 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.

Panel A: Order book variables	
Variable	Description
<i>Ret</i>	Mid-price returns computed from the average of the level-1 bid and ask prices at the end of each time period.
<i>Level</i>	The average of the volume of shares available for immediate execution at levels 1-10 of the order book.
<i>Slope</i>	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.
<i>Curve</i>	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
<i>HFT</i>	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 must be at 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)
<i>Price</i>	The stock-day closing price.
<i>Volume</i>	The daily number of shares sold at the individual stock level.
<i>Return</i>	The daily holding-period return at the individual stock level.
<i>Size</i>	The daily market capitalization at the individual stock level.

Table 2.3. Variable descriptions. 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.

## CHAPTER 3 ORDER BOOK EVENTS ON A POISSON NETWORK

### 3.1 Introduction

Algorithms abound in the modern world, and no less so in the world of modern finance. One of the first victims (or beneficiaries, depending on your perspective) of automation in has been the trading floor. From asset managers to market makers, the day-to-day trading of securities on lit exchanges is performed by machines, with limited human supervision. Automated trading in the finance literature goes back to the order execution algorithms of Almgren and Chriss (2000) and Bertsimas and Lo (1998). Traditional order execution algorithms rely on diffusion processes to model low-frequency price movements and are simplistic in the sense that they only consider the size and timing of market orders as decision variables.

Trading in contemporary exchanges is faster and more complicated than it used to be, requiring researchers and practitioners to rethink their assumptions and address new considerations. Limit orders are widely used by those *supplying* liquidity, as well as those demanding liquidity; order cancellations and replacements, are pervasive; and *order routing* decisions—deciding which exchange to send orders to—matter (Cont and Kukanov (2017)). Recently, researchers have incorporated some of these aspects while considering the optimal trading strategies of market makers (Guilbaud and Pham (2013); Xu (2015)), but the solution methods remain fundamentally unchanged from earlier studies. In particular, the field has preferred to

use dynamic optimization methods (i.e., solving Hamilton-Jacobi-Bellman equations, typically numerically). Machine learning and artificial intelligence have been largely ignored.

Recent advances in the application of machine learning to optimal control offer alternative strategies. Reinforcement learning provides a case in point. Developments in the training of neural networks have demonstrated its potential as a method of solving complex, dynamic optimization problems. For example, reinforcement learning algorithms have achieved world-class performance in a variety of video games (Mnih et al. (2015)) and board games, famously beating the human world champion in Go (Silver et al. (2017)). In the trading context, which takes place at sub-human timescales, similar types of algorithms might provide significant value.

The primary challenge to applying these methods is that they require a data source. Optimization in market microstructure is complicated by the fact that back-testing is infeasible. Unlike investment decisions, we first have to guarantee that any actions an agent performs are consistent with the data: orders have mechanical effects on the state of the order book that are absent in the historical data. There is also the critical problem that we can no longer assume that actions will not affect the market. Sequences of actions will reveal information to the market, leading to orders that are absent from the historical data.

The traditional dynamic optimization approach overcomes this issue by assuming stochastic processes designed to capture features of microstructure considered essential, and which are compatible with numerical methods: one only needs to



calibrate a model from historical data and apply the solver. Reinforcement learning algorithms are “model-free”, but they do not work with historical data—they require a source. The goal of this study is to take the first step towards an improved understanding of how automated trading works in practice by developing a model of limit order book dynamics that can serve as a training environment—a sort of testing lab—for new classes of algorithms attempting to learn optimal trading strategies<sup>1</sup>.

There are two essential features that a simulator intended to train computational agents should possess. First, the simulator should be *responsive*. It is not enough to generate data that matches statistical moments of observed order book data. To learn, an agent needs to be able to interact with the simulator, which requires that it can generate new data based on the agent’s actions. Second, the model needs to be able to capture dynamics at small timescales, as we expect reaction times to be fast in modern markets. A discrete-time model would need to take extremely small time-steps to capture interactions with any degree of accuracy, which would then require datasets that are incredibly sparse. For example, if the timescale necessary is milliseconds (1/1000 of a second), and we observe one event per second, then the data is 99.9% sparse—99.9999% sparse if the necessary timescale is microseconds. Nasdaq message data is recorded with nanosecond precision, meaning that microseconds are well within the measurable accuracy anticipated by the exchange.

---

<sup>1</sup>Nevmyvaka et al. (2006) is a rare example in this vein. They use a simplified version of reinforcement learning, trained on historical data, to learn how to execute a small order. The machine learning literature on high-frequency trading has tended to focus on *prediction* rather than control

The approach I pursue satisfies both criteria by modeling order book events as a system of mutually-exciting point processes in continuous time.

I view order book dynamics as a network in which different types of events (e.g., limit orders at the best offer) influence the probability of events on other “nodes” of the network. These interactions provide a natural way for a simulator to respond to actions of a trading agent, namely by simulating some new data every time the agent acts. The model is purely “excitatory”, meaning that events can increase the intensity of future events, but does not decrease the intensity. Moreover, the model permits a high degree of flexibility regarding the timing of the relationship between events.

### 3.1.1 Summary

I begin by identifying order book events from message data obtained from NASDAQ. Order book events are defined contingent upon on the state of the order book at the time a message arrives. I distinguish twelve order book events based on the side (bid or ask), order type (limit, cancel, or market), and whether or not an order changes either of the best offers. I apply Markov Chain Monte Carlo methods to sample the joint posterior distribution of the parameters in a continuous-time, event-driven model of order book dynamics. The results of the sampling algorithm are used to analyze the dynamic microstructure—and potential simulator features—of a broad selection of stocks. For each stock in the sample, I compute point estimates of the model parameters which I use to analyze the median and cross-section of model

features and performance.

The model captures a variety of general order book characteristics. Several of the order book events display self-exciting behavior, resulting in clustering of these events. Limit orders and cancellations exhibit coupling behavior, especially between limit orders that improve the offer and cancellations that worsen the offer. Market orders arrivals are rare and mostly independent of limit orders and cancellations, but they have a profound influence on the incidence of all types of future events. The model indicates that the timescale of these interactions is small, typically in the range of microseconds to milliseconds—despite the fact that I explicitly permit the model to learn response times ranging up to one minute.

The network model fits the message data better than a baseline model that ignores interactions. On average, the likelihood of the proposed model is more than 4 bits per event higher than the baseline. Through further analysis, I explore which of these features are the most important in explaining order book activity, finding that the majority of this improvement is due to the first 20-25% of the interaction terms. Overall, the inference procedure tends to identify a modest-sized network of events in which interactions between nodes take place on extremely small timescales, or not at all. All of these features are found within a single, consolidated framework that provides a more precise and sophisticated understanding of the interactions between order book events.

### 3.1.2 Contribution

This paper is the first of its kind to analyze a broad cross-section of order books. High-dimensional, continuous-time models of the type employed in this study demand computationally intensive methods when applied to large datasets like those found in market microstructure studies. For these reasons, articles tend to be limited in scope, often restricted to the analysis of a single security. In contrast, this study presents results for a sample of over 400 stocks in the S&P 500, permitting a more general analysis of order book dynamics.

This paper is the first to apply Bayesian inference methods to a continuous-time, event-driven order book model. Traditionally, research in this area has relied on parameterizations that facilitate maximum likelihood estimation. In contrast, following recent modeling approaches lying at the intersection of computational neuroscience and machine learning, I introduce a model and inference procedure that possess greater flexibility in addition to the standard benefits of Bayesian methods.

Several academic articles develop statistical models of market microstructure using modern, high-frequency exchange data. One approach to modeling order flow relies on state variables to reduce the dimension of the order book to a small number of variables that predict changes in prices and volumes (e.g., Avellaneda et al. (2011) and Cont et al. (2013)). This chapter pursues an orthogonal approach, which is to directly model the underlying process of events leading to order book states. Huang et al. (2015) is an example of a compromise between the state-based and event-based approaches in which order arrivals are modeled as inhomogeneous point processes

with intensities that depend on the state of the order book.

This chapter is closely related to Large (2007) who analyzes dynamic microstructure using the same general class of continuous-time model. He focuses on what he calls the “resiliency” of markets, which is the capacity of a market to replace liquidity consumed in a *single* “large” trade. Such a notion made sense for the security analyzed at the time of the study, but its consequence is less evident at present: in lit exchanges, traders split large orders into smaller portions that rarely deplete liquidity beyond what is available at the best offer. I focus on a broader analysis of order book events, including an examination of the cross-section of microstructure dynamics, and introduce a subtle but consequential change to assumptions regarding the timing of interactions between events.

Both papers are generalizations of the analysis of Biais et al. (1995). That study identified patterns in order flow on the Paris Bourse (now Euronext Paris) by examining the conditional probabilities of various order types following a single order, and the expected waiting times between events. The approach taken in the current chapter goes considerably further by estimating a model of order flow that is *fully* conditioned on prior information, and which operates in continuous time. I can, therefore, make detailed statements about the marginal effect of any given order on the probability of future orders, including the precise timing of its impact. In addition, the dataset that I analyze improves significantly over the size and relevancy of the datasets that these two studies used. Although the datasets used in both papers come from fully-automated exchanges, they predate the swell in popularity of

electronic trading and contain orders of magnitude fewer observations.

Finally, the approach taken in this chapter contributes to a broader line of research studying market microstructure through the lens of mutually-exciting point processes. Bowsher (2007) was the first paper to apply such methods to the study of market microstructure. Bacry et al. (2013) models upwards and downwards price movements as mutually-exciting processes to explain patterns in price variance and covariance across timescales. Alfonsi and Blanc (2016) addresses optimal order execution in models where such dynamics drive market orders. Common features to nearly all of this literature are the reliance on exponential decay impulse response functions, low-dimensional models, and maximum likelihood estimators. An exponential impulse response leads to a simplified likelihood expression, and keeping the dimension low (all of the above are bivariate models) reduces computation times, thereby facilitating maximum likelihood estimation. This chapter is the first to my knowledge to propose a Bayesian model and inference procedure. The inference procedure is made possible in the relatively high-dimensional model considered by writing customized Python code that utilizes Cython to overcome computational bottlenecks. The package download and documentation is available at <https://github.com/cswaney/fawkes>.

### 3.2 Model

I model Limit order books as event-driven processes. In choosing a model that can serve as a simulator, we require a continuous-time statistical framework that identifies the exact ordering of events. Also, the model should be *responsive* so

that a user can discover how actions change expected outcomes, rather than merely learning expectations. Point processes—and Poisson processes in particular—provide a statistical framework satisfying both criteria.

### 3.2.1 Mutually-Exciting 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 a constant, then the expected number of events is  $\lambda(\tau_2 - \tau_1)$ , so that  $\lambda$  is the expected frequency of events. In general, the intensity is time-varying (e.g., to capture seasonalities in the frequency of events), in which case the expected number of events is equal to the *integrated intensity*,

$$\mathcal{I}[t_1, t_2] = \int_{\tau_1}^{\tau_2} \lambda(t) dt = \mathbb{E}[N]. \quad (3.1)$$

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 up to that point. Specifically,

$$\lambda(t) = \lambda_0(t) + \sum_{s_m < t} \phi(t - s_m), \quad (3.2)$$

where  $\phi$  is a non-negative *impulse response* function that acts to increase the intensity following each event. In the absence of events, the model functions as a standard Poisson process, but when an event occurs, the model generates a momentary surge in the intensity of event arrivals. This “self-exciting” nature allows the process to model,

for example, gang-related crimes in which an initial crime increases the probability of subsequent crimes because it leads to a series of revenge crimes (Cho et al. (2013)). We say the process is *unstable* if the feedback loop between events causes the intensity to explode in a finite amount of time; otherwise, we consider the process *stable*.

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 observe from a collection of  $K$  Poisson processes the events  $\{\{s_m^{(k)}\}_{m=1}^{M_k}\}_{k=1}^K$ , but that the data appear as 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 individual 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—represented by  $\lambda_0$ —or of a Poisson process initiated by an earlier event whose intensity is given by  $\phi(t - s_m)$  on its support  $[s_m, T]$ . Thus, we can simulate the process according to Algorithm 3.1 below.

A *marked* Poisson process is a Poisson process having multiple event types. Observations of a marked Poisson process consist of event times,  $s_t \in [0, T]$ , and labels,  $c_t \in \{1, \dots, K\}$ . It is natural to think of such a process as a network of processes, in which case each process represents a different *node*. A *mutually-exciting Poisson*



---

**Algorithm 3.1** Generative model for a self-exciting Poisson process

---

```

 $S \leftarrow$  Generate sample  $\{s_m\}_{m=1}^M$  according to  $\lambda(t) = \lambda_0$  on  $[0, T]$ 

 $S_{\text{out}} \leftarrow S$ 

repeat

  for  $s_m$  in  $S$  do

     $S_m \leftarrow$  Generate sample  $\{s_k\}_{k=1}^{M_m}$  according to  $\lambda(t) = \phi(t - s_m)$  on  $[s_m, T]$ 

  end for

   $S = S_1 \cup \dots \cup S_M$ 

   $S_{\text{out}} \leftarrow S_{\text{out}} \cup S$ 

until  $S = \emptyset$ 

```

---

*process* is a marked Poisson process that permits these nodes to communicate with each other in the same manner as in a self-exciting Poisson process. The intensity of the  $n^{\text{th}}$  event type is given by

$$\lambda_n(t) = \lambda_{n,0}(t) + \sum_{s_m < t} W_{c_m,n} \phi_{c_m,n}(t - s_m), \quad (3.3)$$

where  $\phi_{c_m,n}(t - s_m)$  is the non-negative impulse response function governing the interaction between the  $c_m^{\text{th}}$  node and the  $n^{\text{th}}$  node.

For the purposes of simulating a mutually-exciting Poisson process, it is useful to associate each event with a Poisson process having an intensity

$$\lambda_{m,n}(t) = \begin{cases} 0 & t \leq s_m \\ W_{c_m,n} \phi_{c_m,n}(t - s_m) & t > s_m \end{cases} \quad (3.4)$$

Comparing to (3.3), we see that the intensity of each node is a linear combination of a background process,  $\lambda_{n,0}$ , and a sequence of independent Poisson processes. According to the Poisson superposition principle, data generated by this process is indistinguishable from data produced by its component processes. Therefore, we can simulate the process by following Algorithm 3.2 below.

---

**Algorithm 3.2** Generative model for a mutually-exciting Poisson process

---

$$\lambda_{m,n} = W_{c_m,n} \phi_{c_m,n}(t - s_m)$$

$S, C \leftarrow$  Generate sample  $\{s_m, c_m\}_{m=1}^M$  according to  $\lambda(t) = \lambda_0$  on  $[0, T]$

$$S_{\text{out}}, C_{\text{out}} \leftarrow S, C$$

**repeat**

**for**  $s_m, c_m$  in  $S$  **do**

$(S_m, C_m) \leftarrow$  Generate sample  $\{s_k, c_k\}_{k=1}^{M_m}$  according to  $\lambda_{m,n}(t)$  on  $[s_m, T]$

**end for**

$$S, C \leftarrow S_1 \cup \dots \cup S_M, C_1 \cup \dots \cup C_M$$

$$S_{\text{out}} \leftarrow S_{\text{out}} \cup S$$

**until**  $S = \emptyset$

---

### 3.2.2 Order Book Events

Applying the mutually-exciting Poisson model to market microstructure data requires us to define a set of order book events that form the nodes of a graph. In this

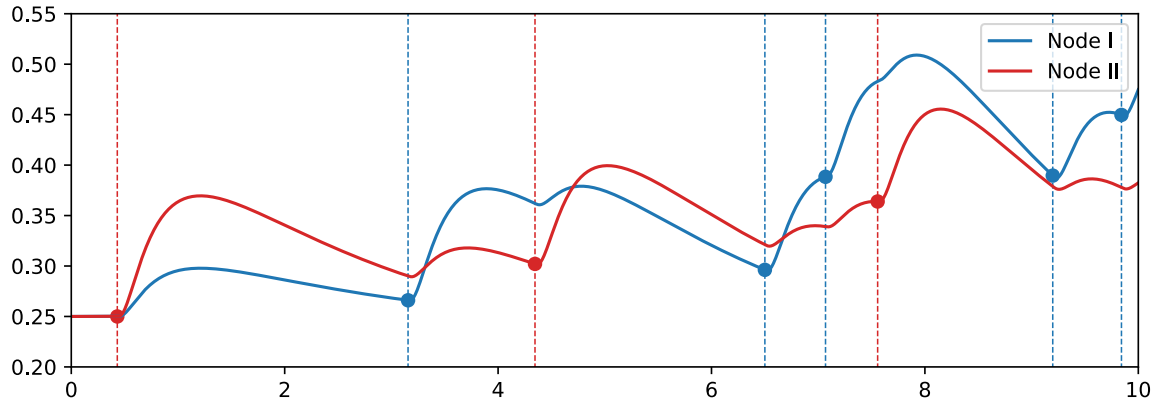


Figure 3.1: Example of a mutually-exciting Poisson process. Event arrivals increase the intensity on both nodes. In the figure, the effect on the event node is twice as strong.

study, I focus on “Level I” of the order book and choose events that affect either the volume or price at the best bid or ask. There are three basic types of orders (limit orders, market orders, and cancellations), which traders submit on either side of the book (bid or ask), and each of these order types either moves or maintains the best offer price. Thus, I define 12 order book events that capture how order flow responds to changes in either the volume or price (or both) at the best offer. These definitions match Large (2007), except that they account for the fact that cancellations also change the best offer when they are the last order remaining at that price.

Table 3.1 describes the order book events. The average market buy and sell orders that fail to move the best offer are for 187 and 181 shares, respectively<sup>2</sup>. Market orders that move the best price are larger on average: 292 shares for buys and 283 shares for sells. These values overstate the size of a typical trade, as the median size

---

<sup>2</sup>Note that the side of a market order event in Table 3.1 indicates the side of the book impacted by the trade, not the side that initiates the trade. Thus, a market order to sell is a market order on the bid side (events 5 and 6)

of market orders is 100 shares for trades that maintain prices, and slightly higher than 100 shares for trades that move prices (111 for buys and 106 for sells). The frequencies of the various event types are similar to the frequencies of the underlying message data shown in Table 3.2. The median number of order book events is slightly greater than one-fifth the median number of messages per day.

Besides ignoring activity below Level I of the book, I also disregard replacement orders and executions against hidden orders. As discussed above, a satisfactory treatment of these events would require additional nodes, and I exclude them in favor of a smaller model. Unlike Biais et al. (1995), I do not distinguish between orders based on their number of shares. Distinctions between orders size are less relevant in modern markets because order sizes have decreased, with the majority of trades involving 100 shares or less and an exponentially decreasing percent distributed across higher numbers of round lots. The event definitions also ignore potentially useful information reflecting the state of the order book. One way to include such information is to make parameters functionally dependent on it. For example, the background intensity parameter could be a linear function of the bid-ask spread or imbalance.

### 3.2.3 Parameterization

I assume that the limit order book network is *fully-connected*, meaning that every event type is allowed to influence every other event type. The mutually-exciting Poisson model permits a variety of alternative assumptions regarding the existence of connections between nodes. For example, links might be determined by the flip

Event #	Order type	Bid or ask?	Changes price?	Count	Frequency	Shares
1	Limit	Bid	No	5,013	19.0	125
2	Limit	Bid	Yes	1,391	5.3	76
3	Limit	Ask	No	4,936	18.7	128
4	Limit	Ask	Yes	1,352	5.1	73
5	Market	Bid	No	517	2.0	181
6	Market	Bid	Yes	467	1.8	283
7	Market	Ask	No	508	1.9	187
8	Market	Ask	Yes	453	1.7	292
9	Cancellation	Bid	No	4,707	17.9	121
10	Cancellation	Bid	Yes	1,199	4.6	62
11	Cancellation	Ask	No	4,688	17.8	124
12	Cancellation	Ask	Yes	1,108	4.2	10

Table 3.1. The table describes the order book events. *Count* is the median number of events per stock-day. *Frequency* is the percent of events of that type per stock-day. *Shares* is the average number of shares per event.

of a coin, or by proximity to a reference location (Linderman (2016)). Assuming a fully-connected network simplifies inference as I only need to assess the strength of interactions between nodes ( $W$ ), which reduces the computational complexity of the Gibbs sampling algorithm described below.

Following Linderman (2016), the impulse response function is specified by a logistic-normal density with parameters  $\theta_{m,n} = \{\mu_{m,n}, \tau_{m,n}\}$ :

$$\begin{aligned}
\phi_{m,n}(\Delta t) &= \frac{1}{Z} \exp \left\{ \frac{-\tau_{m,n}}{2} \left( \sigma^{-1} \left( \frac{\Delta t}{\Delta t_{\max}} \right) - \mu_{m,n} \right)^2 \right\} \\
\sigma^{-1}(x) &= \ln(x/(1-x)) \\
Z &= \frac{\Delta t(\Delta t_{\max} - \Delta t)}{\Delta t_{\max}} \left( \frac{\tau_{m,n}}{2\pi} \right)^{-\frac{1}{2}}
\end{aligned} \tag{3.5}$$

Making  $\phi$  a probability density facilitates interpretation because the connection weights  $W_{m,n}$  become the expected number of events generated on node  $n$  by an event on node

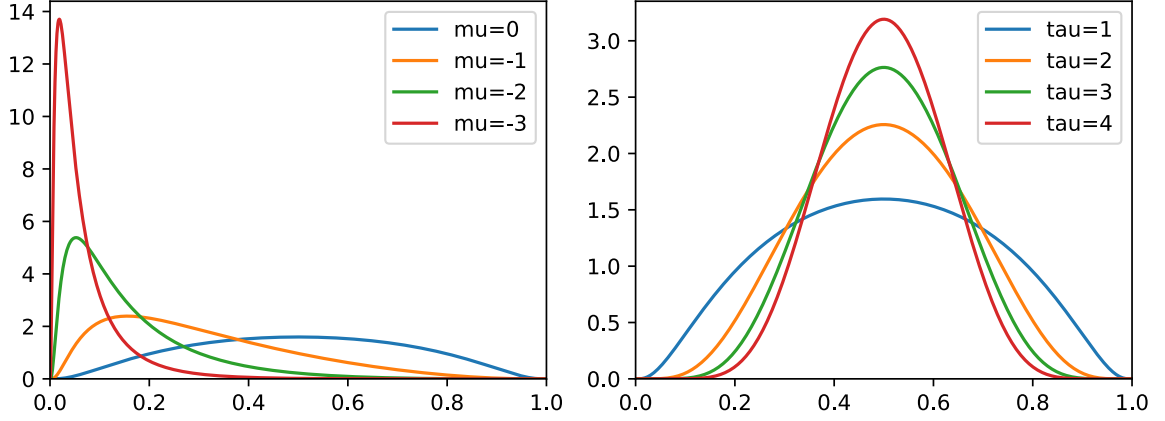


Figure 3.2: Examples of logistic-normal impulse responses. On the left,  $\mu$  determines the apex of the response for fixed  $\tau = 0$ . On the right,  $\tau$  determines the dispersion of the response for fixed  $\mu = 0$ . In both cases,  $\Delta t_{\max} = 1$ .

$m$ . Note that the impulse response is non-zero on the interval  $\Delta t \in [0, \Delta t_{\max}]$ . Thus, events on node  $m$  have no influence over events on node  $n$  after a maximum lag of  $\Delta t_{\max}$  seconds. Finally, as a two-parameter density,  $\phi$  can model a variety of impulse response functions, as demonstrated by Figure 3.2. Importantly, observe that the first parameter ( $\mu$ ) controls the apex of the response, while the second ( $\tau$ ) determines its dispersion.

The maximum lag  $\Delta t_{\max}$  is set to 60 seconds. Based on the results in the next section, this is longer than required (and causes the estimation to be much slower in practice), but allows for the possibility that events have important low-frequency consequences. It is also possible that events affect dynamics over multiple timescales. The same inference procedure is easily modified so that some or all of the events have both fast (small  $\Delta t_{\max}$ ) and slow (large  $\Delta t_{\max}$ ) responses.

Finally, I assume that the background intensity of events is constant:  $\lambda_0(t) =$

$\lambda$ . As previously remarked, overall market activity displays a distinct ‘U’-shape during the trading hours from 9:30 to 16:00, and I accordingly restrict the sample to the relatively stable trading hours between 10:30 and 15:00. Alternatively, a piecewise constant or linear spline could account for seasonalities and other un-modeled exogenous variation, or  $\lambda_0$  could be allowed to be a function of additional covariates (e.g., news events).

### 3.2.4 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 based on work in the field of computational neuroscience (Linderman (2016)). The algorithm uses standard Gibbs sampling to generate approximate samples from the conditional posterior distributions of the model parameters using conjugate priors, as described below.

#### Weights

If the weights have gamma priors,

$$w_{n,n'} \sim \Gamma(\kappa, \nu_{n,n'}), \quad (3.6)$$

then their conditional distributions are also gamma:

$$p(w_{n,n'} | \{s_m, c_m, \omega_m\}_{m=1}^M, a_{n,n'} = 1, \kappa, \nu_{n,n'}) = \Gamma(\tilde{\kappa}_{n,n'}, \tilde{\nu}_{n,n'}), \quad (3.7)$$

where

$$\tilde{\kappa}_{n,n'} = \kappa + M_{n,n'}, \quad (3.8)$$

$$\tilde{\nu}_{n,n'} = \nu_{n,n'} + M_n, \quad (3.9)$$

and

$$M_n = \sum_{m=1}^M \mathbb{I}[c_m = n], \quad (3.10)$$

$$M_{n,n'} = \sum_{m=1}^M \mathbb{I}[c_m = n] \mathbb{I}[c_{m'} = n'] \mathbb{I}[\omega_{m'} = m]. \quad (3.11)$$

The first sufficient statistic  $M_n$  is the number events occuring on node  $n$ . The sufficient statistic  $M_{n,n'}$  is the number of events on node  $n'$  generated by events on node  $n$ .

### Impulse Response

The likelihood is conjugate with a normal-gamma distribution. If the priors of the impulse response parameters are normal-gamma,

$$(\mu_{n,n'}, \tau_{n,n'}) \sim \mathcal{NG}(\mu_\mu, \kappa_\mu, \alpha_\tau, \beta_\tau), \quad (3.12)$$

then the conditional posterior distributions are normal-gamma with parameters

$$\tilde{\mu}_{n,n'} = \frac{\kappa_\mu \mu_\mu + M_{n,n'} \bar{x}_{n,n'}}{\kappa_\mu + M_{n,n'}}, \quad (3.13)$$

$$\tilde{\kappa}_{n,n'} = \kappa_\mu + M_{n,n'}, \quad (3.14)$$

$$\tilde{\alpha}_{n,n'} = \alpha_\tau + \frac{M_{n,n'}}{2}, \quad (3.15)$$

$$\tilde{\beta}_{n,n'} = \frac{\nu_{n,n'}}{2} + \frac{M_{n,n'} \kappa_\mu (\bar{x}_{n,n'} - \mu_\mu)^2}{2(M_{n,n'} + \kappa_\mu)}, \quad (3.16)$$

The sufficients statistics in this case are defined with respect to

$$x_{m,m'} = \ln \left( \frac{s_{m'} - s_m}{\Delta t_{max} - (s_{m'} - s_m)} \right), \quad (3.17)$$



which is the log of the ratio of the time elapsed since event  $m$  occurred to the time remaining until event  $m$  can no longer generate events. The sufficient statistics are the mean and variance of  $x_{m,m'}$ :

$$\bar{x}_{n,n'} = \frac{1}{M_{n,n'}} \sum_m \sum_{m'} \mathbb{I}[c_m = n] \mathbb{I}[c_{m'} = n'] \mathbb{I}[\omega_{m'} = m] x_{m,m'}, \quad (3.18)$$

$$\bar{\nu}_{n,n'} = \frac{1}{M_{n,n'}} \sum_m \sum_{m'} \mathbb{I}[c_m = n] \mathbb{I}[c_{m'} = n'] \mathbb{I}[\omega_{m'} = m] (x_{m,m'} - \bar{x}_{n,n'})^2 \quad (3.19)$$

## Biases

If the background rates have gamma priors,

$$\lambda_n^0 \sim \Gamma(\alpha_n^0, \beta_n^0), \quad (3.20)$$

then their conditional distributions are also gamma:

$$p(\lambda_n^0 | \{s_m, c_m, \omega_m\}_{m=1}^M, \alpha_n^0, \beta_n^0) \sim \Gamma(\tilde{\alpha}_n^0, \tilde{\beta}_n^0), \quad (3.21)$$

where

$$\tilde{\alpha}_n^0 = \alpha_n^0 + \sum_{m=1}^M \mathbb{I}(c_m = n) \mathbb{I}(\omega_m = 0) = \alpha_n^0 + M_n^0, \quad (3.22)$$

$$\tilde{\beta}_0^n = \beta_0 + T. \quad (3.23)$$

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

## Parents

In order to sample the above model parameters, we first need to sample the latent “parent” variables. Their conditional distribution is a multinomial with

probability mass function

$$p(\omega_m = n \mid s_m, \{\lambda_n(t)\}_{n=1}^M) = \frac{\lambda_n(s_m)}{\lambda_{c_m,0} + \sum_{n'=1}^{m-1} \lambda_{n'}(s_m)}, \quad (3.24)$$

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

$$\lambda_{n'}(s_m) = w_{c_{n'}, c_m} \cdot \phi(s_m - s_{n'}; \theta_{c_m, c_{n'}}) \quad (3.25)$$

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 on the  $c_m^{th}$  node at the time of  $m^{th}$  event. As a result, the most likely parent at any time is not, in general, the most recent event, but the event that is most active at that time. Events can also be caused by the background process, which is reflected in the inclusion of  $n = 0$  in equation (3.24). Note that the conditional distributions are independent. This feature allows parents to be sampled in parallel, which substantially reduces the computation time.

## Likelihood

In the results section I compare the network model with a baseline model. I base my comparison on the likelihood of the data given each of the models. The likelihood of observations  $(s, c) = \{s_m, c_m\}_{m=1}^M$  for the network poisson model with parameters estimates  $\theta = (\lambda_0, W, \mu, \tau)$  is given by

$$\begin{aligned} p(s, c \mid \hat{\theta}) &= \prod_{n=1}^N \mathcal{PP} \left( \lambda_n^{(0)}(t) + \sum_{m=1}^M \phi_{m,n}(t) \right) \\ &= \prod_{n=1}^N \exp - \left( \int_0^T \lambda_n^{(0)}(t) + \sum_{m=1}^M \phi_{m,n}(t) \, dt \right) \end{aligned}$$

$$\begin{aligned}
& \times \prod_{m'=1}^M \left( \lambda_n^{(0)}(s_{m'}) + \sum_{m=1}^M \phi_{m,n}(s_{m'}) \right)^{\mathbb{I}(c_{m'}=n)} \\
& = \prod_{n=1}^N \exp - \left( T \lambda_n^{(0)} + \sum_{m=1}^M W_{m,n} \right) \prod_{c_{m'}=n} \left( \lambda_n^{(0)} + \sum_{m=1}^M \phi_{m,n}(s_{m'}) \right) \quad (3.26)
\end{aligned}$$

### 3.3 Data

This study relies on event data for 401 stocks in the S&P 500 on July 24, 2013. The data comes from the Nasdaq HistoricalView-ITCH database, which consists of daily message files recording every update to the limit order book for all securities traded on the Nasdaq exchange with nanosecond precision. I used the message data to reconstruct the limit order book of each stock and identified events based on the state of the order book immediately before their arrival. Raw messages were processed using custom Python code available from <https://github.com/cswaney/hfttools>.

ITCH message data consists primarily of *add*, *delete*, and *execute* messages. Add messages record new limit orders; delete messages indicate cancellations of resting limit orders; execute messages correspond to market orders (or “marketable” limit orders). If an order matches against multiple resting limit orders, then it generates a sequence of execute messages with identical timestamps. I identify such trades with the last execute message reported. If the complete trade moves the best offer, then it is identified with an execute message with a different price from the first message in the sequence. Given the nanosecond precision of ITCH timestamps, I register all other execute messages as separate trades.

Table 3.2 provides descriptive statistics for the ITCH message data. Including

	Frequency	Mean	Min	$q_{0.25}$	Median	$q_{0.75}$	Max
A	47.2	61,769	4,495	42,441	59,551	77,400	163,362
C	0.1	72	1	23	52	101	380
D	45.6	59,706	4,515	41,320	58,330	74,438	161,476
E	2.5	3,264	106	1,683	2,763	4,394	31,713
U	3.7	4,897	21	1,867	3,950	6,189	36,076
X	0.3	405	1	16	107	468	11,809

Table 3.2. Summary statistics for ITCH message data. The table describes the distribution of message counts across the full sample. *Frequency* is the percent of all messages of that type. All other values are computed are numbers of observations per stock-day.

replace messages, add and delete messages constitute 97.1% of all messages: 47.2% of messages are new limit orders, while 45.6% are limit order cancellations. Thus, traders cancel 97% of limit orders before they find a match, which reflects the activity of high-frequency market makers competing for order flow while attempting to avoid being adversely selected. Execute messages—which overstate the number of *trades*—amount to just 2.8% of all messages. While the average stock received over 200,000 messages per day, the majority of this activity is not directly utilized by this study because it does not affect the top of the order book.

ITCH includes additional message types that I ignore in this study. Executions against hidden orders constitute 0.3% of all messages, and 10% off all execute messages. Non-displayed orders typically execute between the bid and ask, in which case they fall outside the event definitions above. It is possible that a hidden order matches at the best offer, but this only happens when a market order specifying more shares than what is displayed at the current best offer hits the book at the same time that a non-displayed order resides at the back of the queue. This outcome is

equivalent to a market order that moves the *displayed* price, which is already defined as an order book event.

I also disregard *replace* (3.7%) and *cancel* (0.1%) messages. Traders use replace orders to change the price of open orders. A replace order is equivalent to simultaneous add and delete messages, but recording both events is inconsistent with the point process specification, and adding additional event types is undesirable. Cancellation messages occur when a trader cancels a fraction of an order.

Figure 3.3 demonstrates patterns in the timing of order book activity in my sample. The unconditional arrival rates of add, delete and execute messages display the typical ‘U’-shape of order book activity, with higher rates at the opening and close of trading hours. This study only looks at events that occur during the relatively stable trading hours between 10:30 and 15:00. The figure also demonstrates that orders are not uniformly distributed throughout each second. A disproportionate number of orders arrive within the first few milliseconds of each second. As noted by Hasbrouck and Saar (2013), this feature is probably due to algorithms that submit orders at even intervals of time.

### 3.4 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 de-

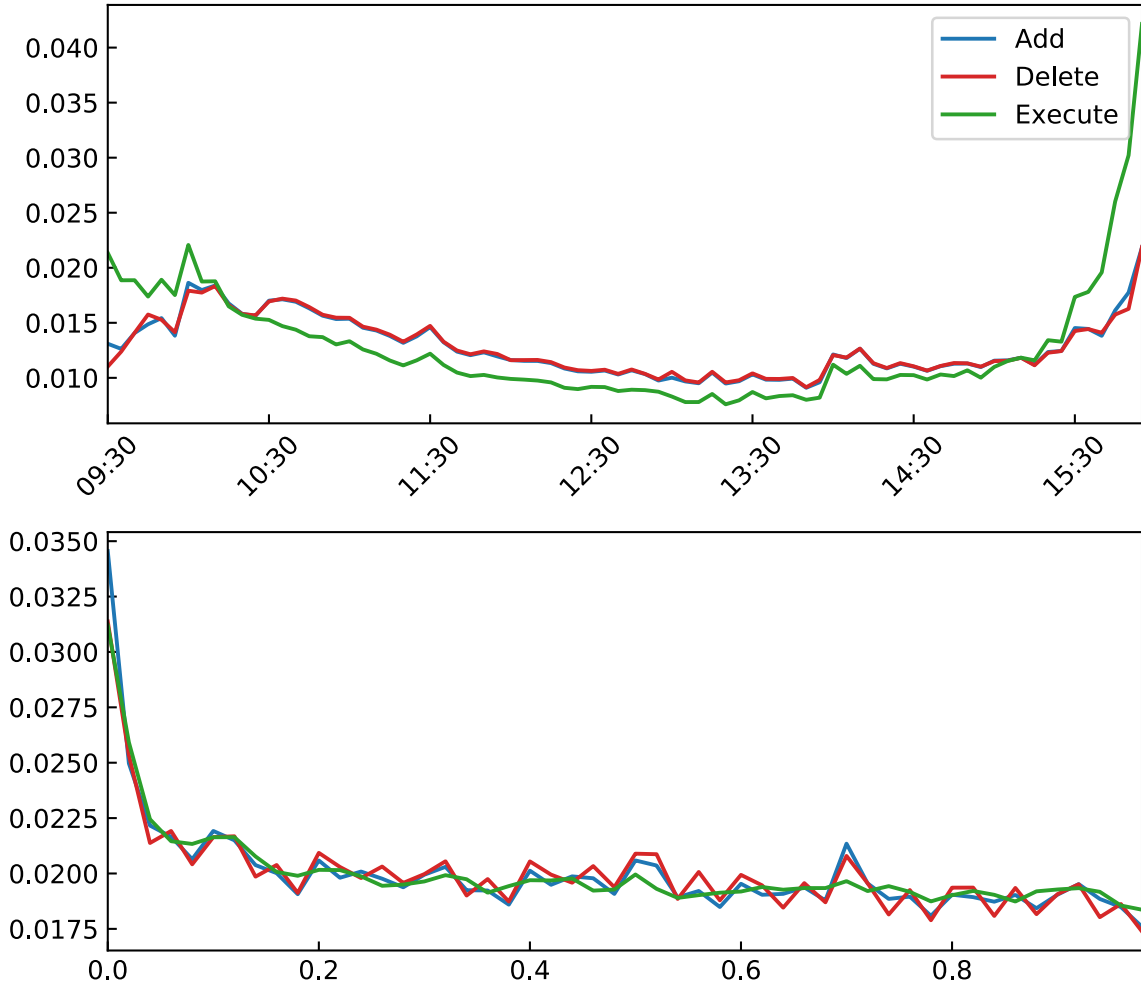


Figure 3.3: Patterns in order book message activity. The figure describes intra-day seasonalities in the arrival of messages averaged across the full sample. The top figure shows the distribution of *add*, *delete*, and *execute* messages within each trading day. The bottom figure shows the distribution of messages within each second.

scribed in previous sections using a single day of message data. The point estimator considered is the posterior median of the Gibbs sample, which minimizes posterior absolute loss risk. As the MCMC algorithm employed scales on the order of the number of observations squared, I reduce the total sampling time by excluding stocks in the S&P 500 with more than 60,000 events between 10:30 am and 3:00 pm, which

leaves a total of 401 stocks.

### 3.4.1 Connections

What types of interactions characterize aggregate limit order book dynamics? Figure 3.4 presents the median point estimate of the connection weight matrix  $W$  across stocks. I begin by considering the raw point estimates shown on the left. The rows of this figure are labeled by parent event and the columns are labeled by child event. Reading across the rows of the matrix allows us to see the impact of that row event on all other event intensities; reading down the columns allows us to see how much each event impacts the intensity of the column event.

There are a few general observations to be made. First, the median point estimates are symmetric across sides of the orders book. If there is a pronounced connection between an order type on the bid side of the book to an order type on the ask side of the book, then we will also observe a connection of similar magnitude between the same order types, but from the ask to the bid side. Second, all of the order types (limit, market and cancel) display self-exciting behavior, which can be seen by looking at the diagonal of the matrix. Third, connection weights are typically larger than background rates, with the median point estimate of  $W_{m,n}$  being as much as 120 times  $\lambda_{n,0}$ . This observation alone suggests that the mutually-exciting feature plays a central role in modeling order flow.

The top four rows of the matrix tell us how limit orders influence subsequent order flow. Limit orders display self-exciting behavior, but the effect is absent for

limit orders that improve the offer. Instead, these orders are followed by (same-side) limit orders at the new offer. In fact, these follow-on orders are the primary self-connections between limit orders, demonstrating that price-improving orders are significant order book events that traders respond to by supplying fresh liquidity to preserve execution priority. Limit orders at the offer precede (same side) cancellations at the offer, while limit orders that improve the offer are linked with subsequent (same side) cancellations, with little difference seen between cancellations that worsen the offer and those at the offer. There are no connections between limit orders and future market orders.

The bottom four rows of the matrix tell us how order cancellations influence subsequent order flow. Connections from order cancellations to following order cancellations display nearly the same pattern as the self-connections between limit orders. There appears to be mild clustering of order cancellations at the offer, but the stronger effect is between cancellations that worsen the offer and following (same side) cancellations at the offer. Order cancellations that alter the offer are followed by (same side) limit orders that improve the offer. This observation is a counterpart to the opposite effect seen for limit orders above. Taken together, these effects form a cycle of “fleeting liquidity”. As with limit orders, there are no connections between order cancellations and future market orders.

The middle four rows of the matrix tell us how market orders influence subsequent order flow. While we have seen that limit orders and cancellations do not affect market orders, market orders are critical influencers of both events. Market orders



display self-excitation (clustering) between trades at the offer, but not between trades that worsen the offer. The absence of the latter self-connection is in some sense mechanical: the median trades is small, so even if a sequence of orders matches all orders at the top of the book, the trade that finally exhausts liquidity will be followed by a market order at the offer. However, observe that the latter is absent from the results, suggesting that market orders are submitted until liquidity is exhausted, and then traders wait to see if liquidity will be replaced, or pause to conceal their intentions.

The most significant effects in the raw connection weight matrix are between market orders that worsen the offer and subsequent limit orders on the opposite side of the book, particularly at the offer. The model predicts trades that exhaust liquidity at the offer are most likely to generate opposite-side limit orders at the offer, opposite-side limit orders that improve the offer, or same-side limit orders that better the offer, in that order. The reaction to market orders at the offer is relatively unexciting.

Market orders display a strong connection to subsequent cancellations. Market orders that worsen the spread are followed by cancellations at the offer on either side of the book. This effect is interesting because it is at odds with the relationship between market orders and limit orders. Cancellations on the opposite side of the book are consistent with traders replacing orders at an improved offer. There appears to be a cautious response (cancellations) and an aggressive response (limit orders) to these trades.

Next, we turn our attention to the *normalized* weight matrix on the right of Figure 3.4. In this figure, the median point estimates of  $W_{m,n}$  are divided by

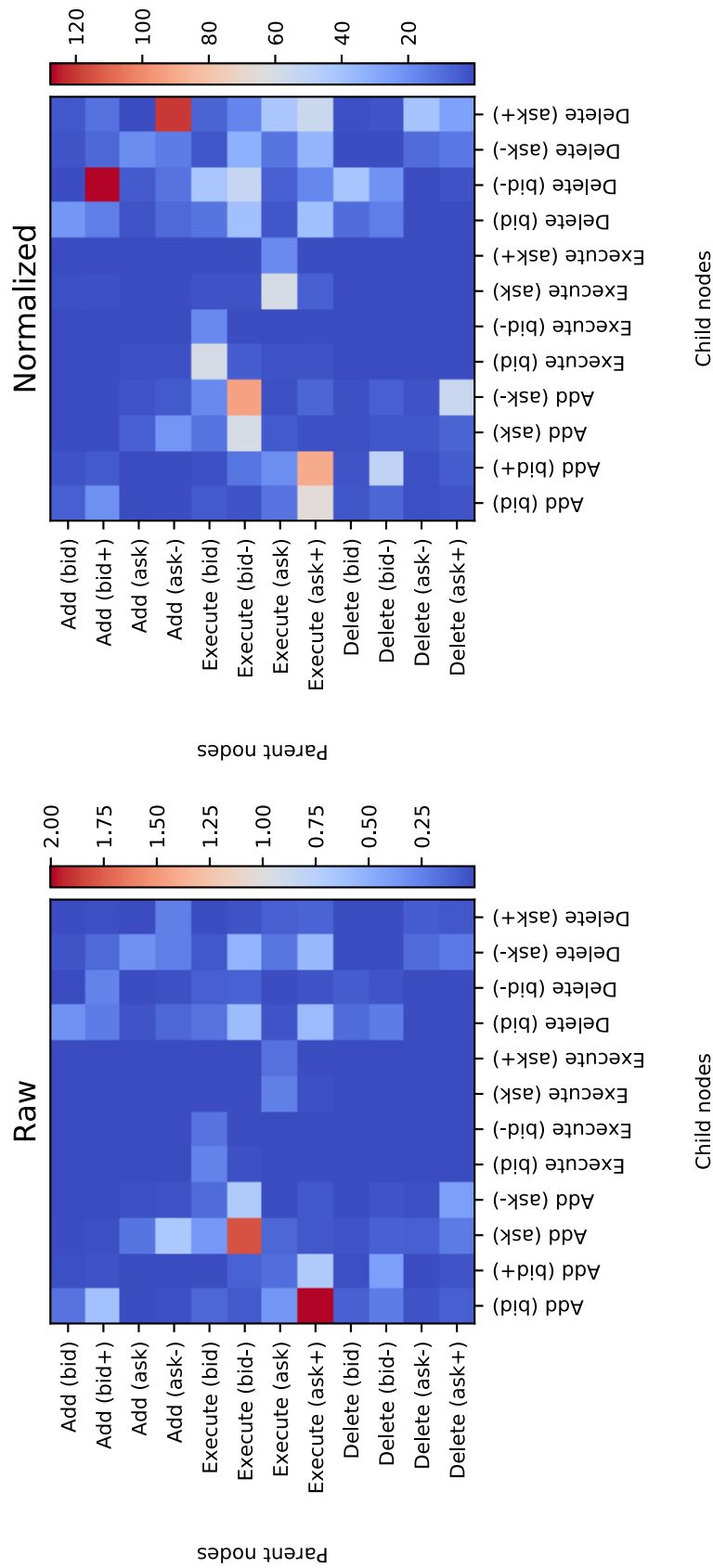


Figure 3.4. Median connection weight point estimates. The figure shows the median connection strength between event types across the full sample of stocks. Larger values indicate stronger connections, which are expected to generate a greater number of child events. (Left) The median of connection weight 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.

the median point estimates of  $\lambda_{0,n}$ . The resulting values are therefore the relative impacts on the child node (the units are expected events divided by expected events per second). This view allows us to explore which events have the biggest effects at the event level and accounts for the fact that some events are more frequent than others. These are not necessarily the most important connections from the view of explaining the data as a whole.

The most significant change is that the connection between limit orders that improve the offer and cancellations that worsen the offer is magnified under the relative perspective. In fact, they are the dominant connections in the model viewed in this manner. 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. 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. 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 per second as background activity.

The relative view also highlights the connection between cancellations that worsen the offer and limit orders that improve the offer so that the complete fleeting order (or “flickering” quotes) cycle is particularly evident. The relative viewpoint also further highlights the connections between market orders and limit orders, except that market orders that worsen the offer now have a larger relative impact on limit

orders that improve the offer than on limit orders that do not. Another pair of connections that appears much more clearly under this normalization is that between executions and subsequent executions. After normalization, observe the magnitude of the clustering effect more clearly. Trades are expected to generate between 40 and 60 times as many trades as we would expect to occur in one second due to exogenous supply and demand alone, depending on whether the trades exhausts liquidity at the offer or not.

Overall, the most significant change seen by switching to the relative view is on connections to limit orders and cancellations that improve and worsen the offer. The raw weights on these events have similar magnitudes as their counterparts that do not change the offer, but they occur about a quarter as often. The relative view does not merely exaggerate the weights for rare events, however: we observe trades less frequently than any other event, but there is no qualitative difference between the raw and relative views of the impacts of limit order and cancellations on their intensity.

### 3.4.2 Impulse Response

The model permits a detailed view of the timescale of interactions between event nodes via the impulse responses. I define the median impulse response as the impulse response function evaluated at the posterior median point estimate,  $\hat{\phi}_{m,n}(t) = \phi_{m,n}(t \mid \hat{\mu}_{m,n}, \hat{\tau}_{m,n})$ . There is substantial variation in the timing of reactions. The response self-exciting response of market orders is nearly instantaneous, whereas the

weak connections from limit orders and cancellations to market orders operate at timescales on the order of seconds.

Figure 3.5 plots the median impulse responses of four pairs of connections discussed above. I plot the responses in log time to facilitate interpretation (in logarithmic time, the response is a normal distribution). The plots on the left show the response to market orders that worsen the offer of limit orders at the offer (above) and limit orders that improve the offer (below). Both reactions peak between one microsecond ( $10^{-6}$  seconds) and one millisecond ( $10^{-3}$  seconds), where their probability-mass concentrates, and of the four connections shown, these are the fastest. It is interesting to note that the response by limit orders that improve the offer is faster than the response of limit order at the offer, consistent with offers queueing behind a newly established best price.

The right side of Figure 3.5 shows the response of cancellations to market orders that worsen the offer (above) and to limit orders that improve the offer (below). The impulse response of deletions to market orders attains a maximum intensity around one millisecond, while the intensity of cancellation orders to limit orders reaches a peak between one millisecond and one second. Thus, flickering orders are in fact permitted to sit for a relatively extended period before being canceled. The reverse response of limit orders following cancellations (not shown) occurs nearly instantaneously, suggesting that such events are, effectively, replace orders.

It is worth contrasting these results with exponential impulse response functions. An exponential impulse response function creates an immediate increase in the

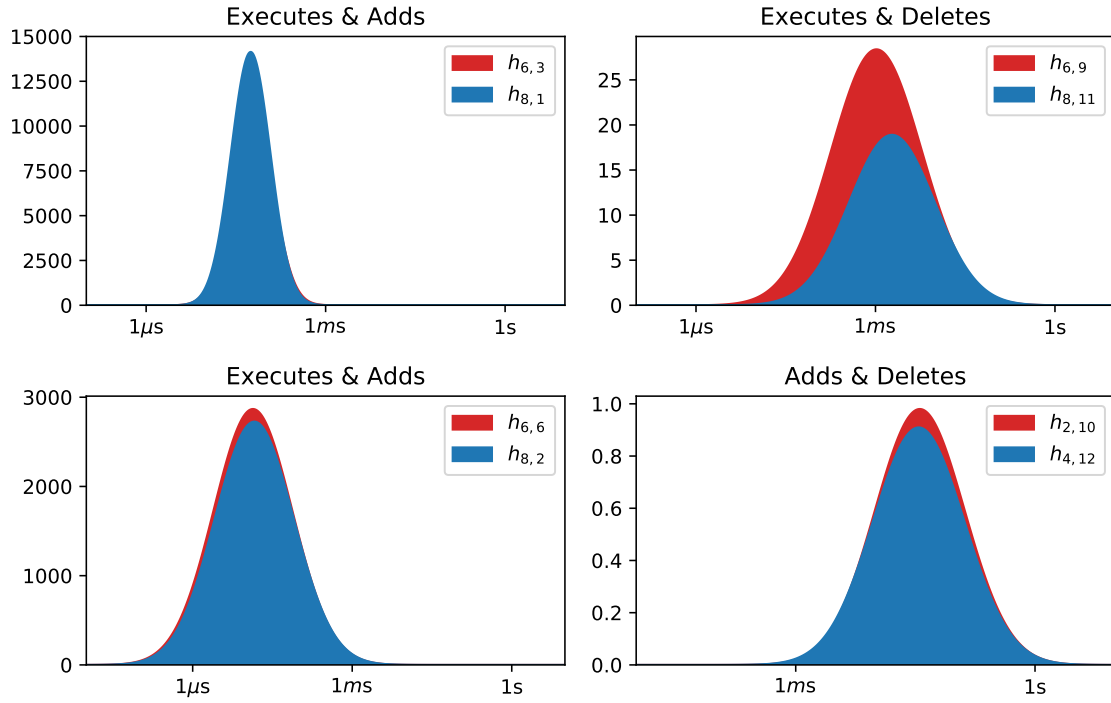


Figure 3.5. Median impulse response point estimates. The figure shows the median impulse response for selected parent-child event connection pairs across the full sample. The horizontal axis is logarithmic time to improve interoperability. The figure demonstrates that important connections operate at 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.

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 a child event intensity spikes. Measured at timescales on the order of minutes (or even seconds), the distinction appears unimportant. As these results demonstrate, interactions between order book events take place on timescales measured in milliseconds, and on this timescale the difference is meaningful.

Figure 3.6 gives a different sense of the character of the impulse response by showing an example of a fitted intensity for SIG. The intensity is estimated using the posterior median of the model parameters. The top panel shows the estimated intensity of market sell orders that decrease the best bid between 10:30 and 11:00, and the bottom panel display a scatterplot of all events during the same period with market sell events highlighted. At this scale, the fitted intensity appears as a sequence of “spikes” that are principally induced by message activity off of the highlighted node. The spikes sometimes—but not always—precede events on this channel. All of the large spikes between 10:30 and 10:40 are followed by market sell orders, but the isolated spike around 10:48, which appears at the end of an intense cluster of activity, is not immediately followed by an event.

### 3.4.3 Further Analysis

The features identified by the model thus far are intriguing, but do they explain the data well? Next, I briefly examine the issues of model fit and stability.

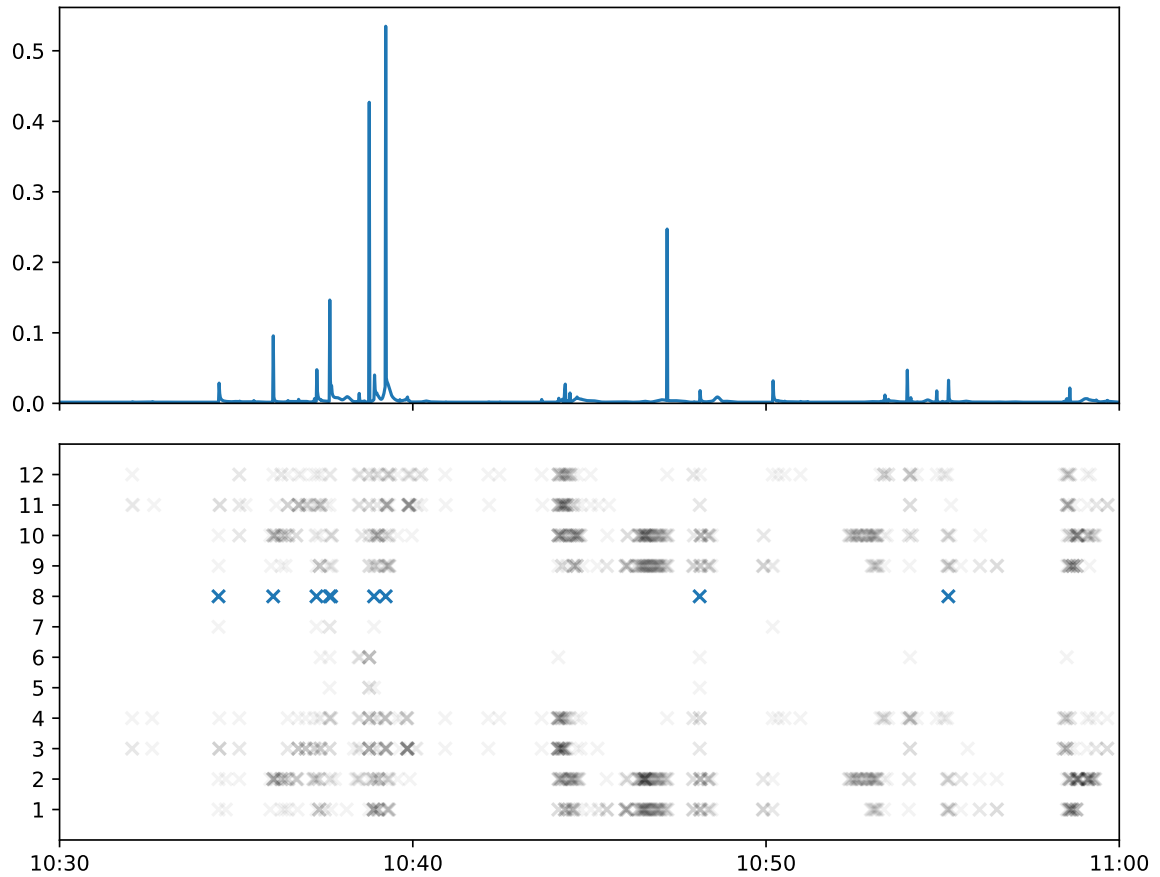


Figure 3.6: Example of a daily fitted intensity: SIG. The top figure shows the estimated intensity of market buy orders that move the best ask from 10:30 to 11:00 on July 24, 2013 for SIG. The bottom figure indicates event arrivals on each of the network nodes during the same period.



### 3.4.3.1 Likelihood

Figure 3.7 compares the fit of the network model versus the baseline model. I take log-likelihood (calculated in (3.26)) as a measure of model fit. In particular, I compare the log-likelihood of the network model estimates to a standard (homogeneous) Poisson process with a constant intensity vector, which serves as a baseline model. For each stock, the point estimate of the network model is the approximate posterior median of the Gibbs sample described above. Setting the weight matrix to zero and computing the approximate median of the posterior distribution by direct sampling produces a Bayesian point estimate of the baseline model. I calculate the values in the histogram 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 7.0. These results quantify the extent to which a complex interaction of events determines order book dynamics.

The wide range of values in Figure 3.7 demonstrates that the network model fits some stocks better than others. Figure 3.8 compares the fit of the model to characteristics of the limit order book. Specifically, for each stock, I use the same message data used to identify order book events to calculate four order book attributes. The

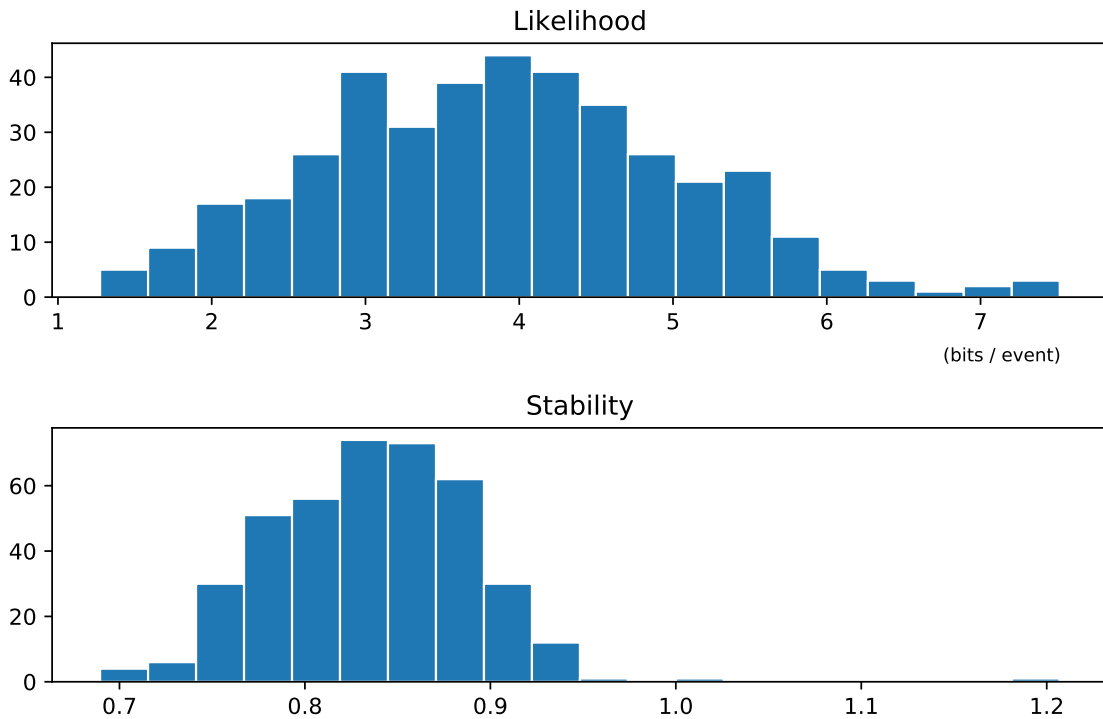


Figure 3.7. Distributions of likelihood & stability. The figure shows histograms of the likelihood and stability of model estimates across the full sample. To compute the likelihood, I select random subsamples from the full sample of daily events on which the model was estimated and average the likelihood 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.

characteristics are defined as follows: *Volume* is the total number of shares traded as determined 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 3.8 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.

### 3.4.3.2 Stability

Confirming model stability is essential because the network model is only physically meaningful (and useful as a simulator) if it is stable. The maximum of the absolute values of the eigenvalues of the connection matrix determines the stability of the model. Models with a maximum eigenvalue greater than one are unstable, meaning that with positive probability, they generate an infinite number of events in a finite amount of time. At the other extreme, a model with a maximum eigenvalue of zero is a homogeneous Poisson process.

Figure 3.7 presents a histogram of these eigenvalues based on median point estimates. The median value is approximately 0.85, which is well below both the instability threshold. Only 2 of the 401 the models have unstable dynamics according to this measure. Figure 3.9 shows scatterplots of stability versus the microstructure characteristics from Figure 3.8. Stability appears to decrease with lower trading vol-

ume 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.

### 3.4.3.3 Model Selection

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 3.10 and 3.11 demonstrate the results for two randomly selected stocks (LLY and SIG). The plots display similar patterns: including weights whose logarithm is less than -4—approximately half of 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 3.4. Figure 3.12 recreates

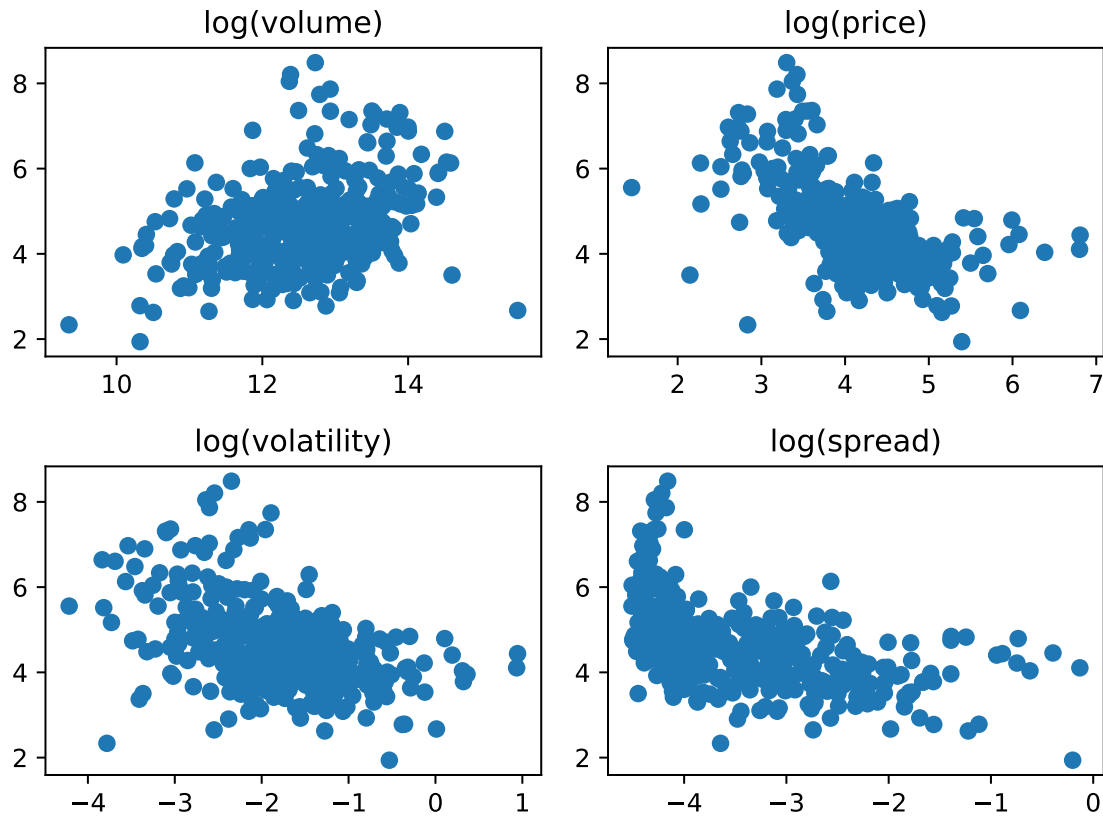


Figure 3.8. Likelihood versus order book characteristics. The figure shows scatterplots of the likelihood of estimated models versus properties of the order books calculated from the message data used to identified order book events (i.e., the same day and hours). Volume is the total number of shares 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.

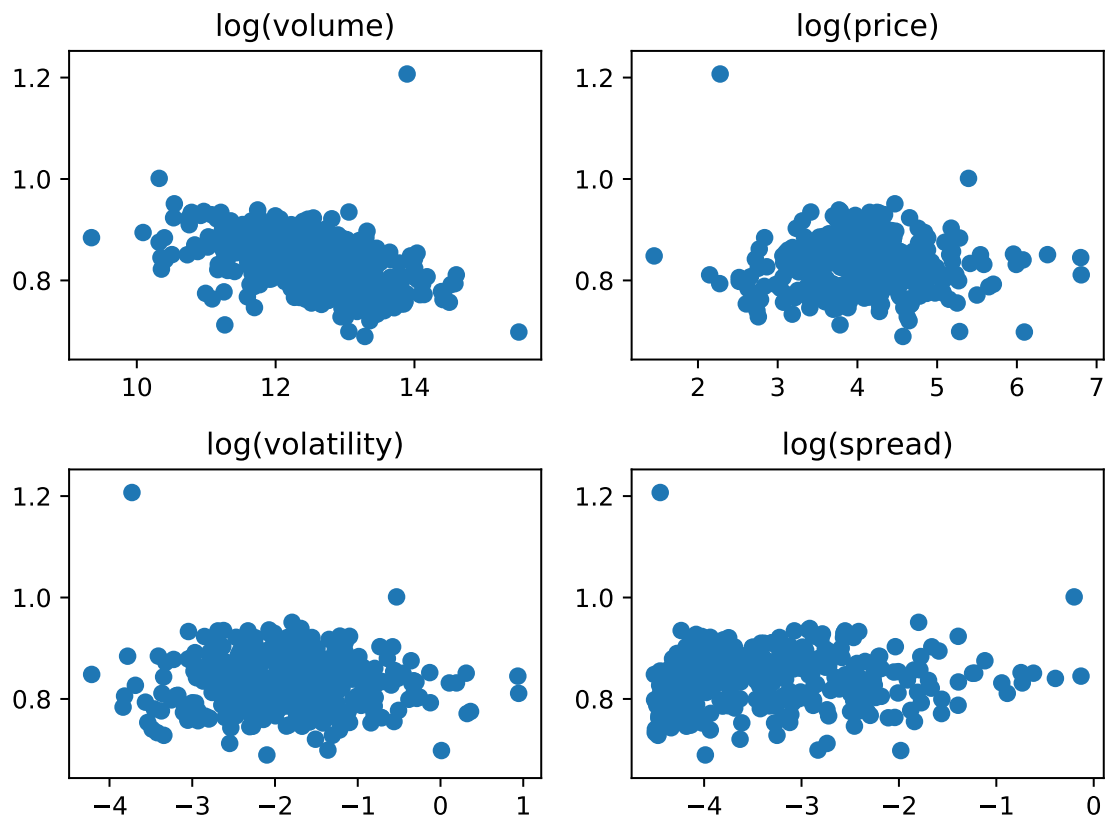


Figure 3.9. Stability versus order book characteristics. The figure shows scatterplots of the stability of estimated models versus properties of the order books calculated from the message data used to identified order book events (i.e., the same day and hours). Volume is the total number of shares traded, 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.

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.

### 3.5 Conclusion

This chapter set out to develop an event-driven model of limit order book dynamics that might serve as the basis of a limit order book simulator. I adopt a mutually-exciting Poisson process model and analyze the model by applying a Bayesian inference procedure to a broad cross-section of stocks in the S&P 500. The model simultaneously captures a variety of essential connections between order book events at the top of the book, and uniquely reveals the timescale of these interactions. The proposed model is stable and leads to sizeable improvements in model fit relative to a baseline Poisson model. Most of the benefits of the model appear to derive from the top 20-25% of model connections. The model permits numerous possible future improvements and additions that might reveal additional features of order book dynamics and enhance its usefulness as the basis of an order book simulator.

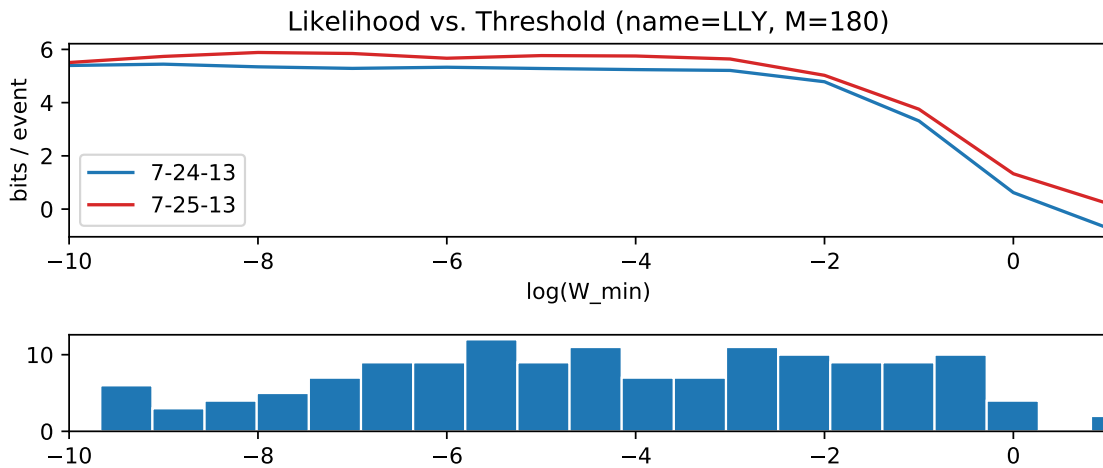


Figure 3.10: Likelihood versus connection weight threshold: LLY. The figure shows the improvement in the likelihood of the network model compared to a standard Poisson model for various thresholds levels. (Top) For each threshold level, the likelihood improvement is computed using the estimated connection 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.



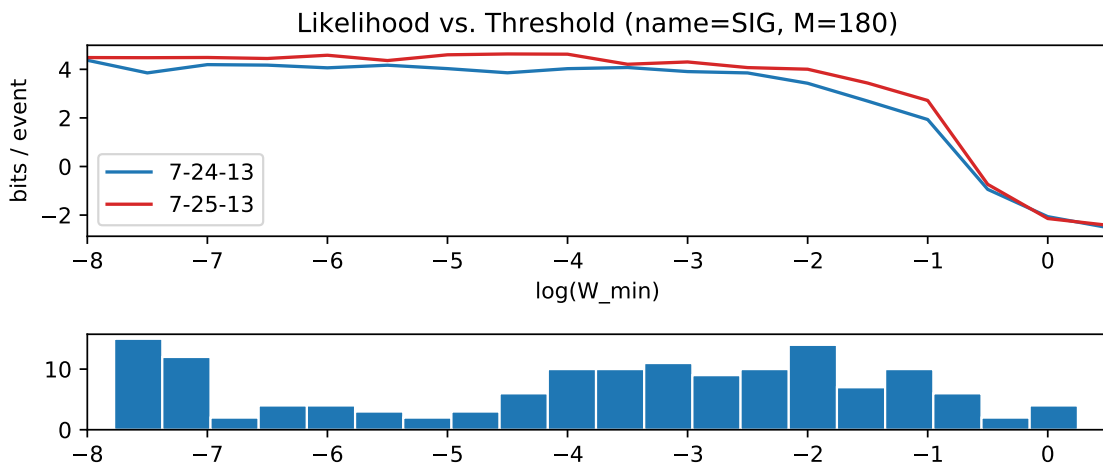


Figure 3.11: Likelihood versus connection weight threshold: SIG. The figure shows the improvement in the likelihood of the network model compared to a standard Poisson model for various thresholds levels. (Top) For each threshold level, the likelihood improvement is computed using the estimated connection 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.

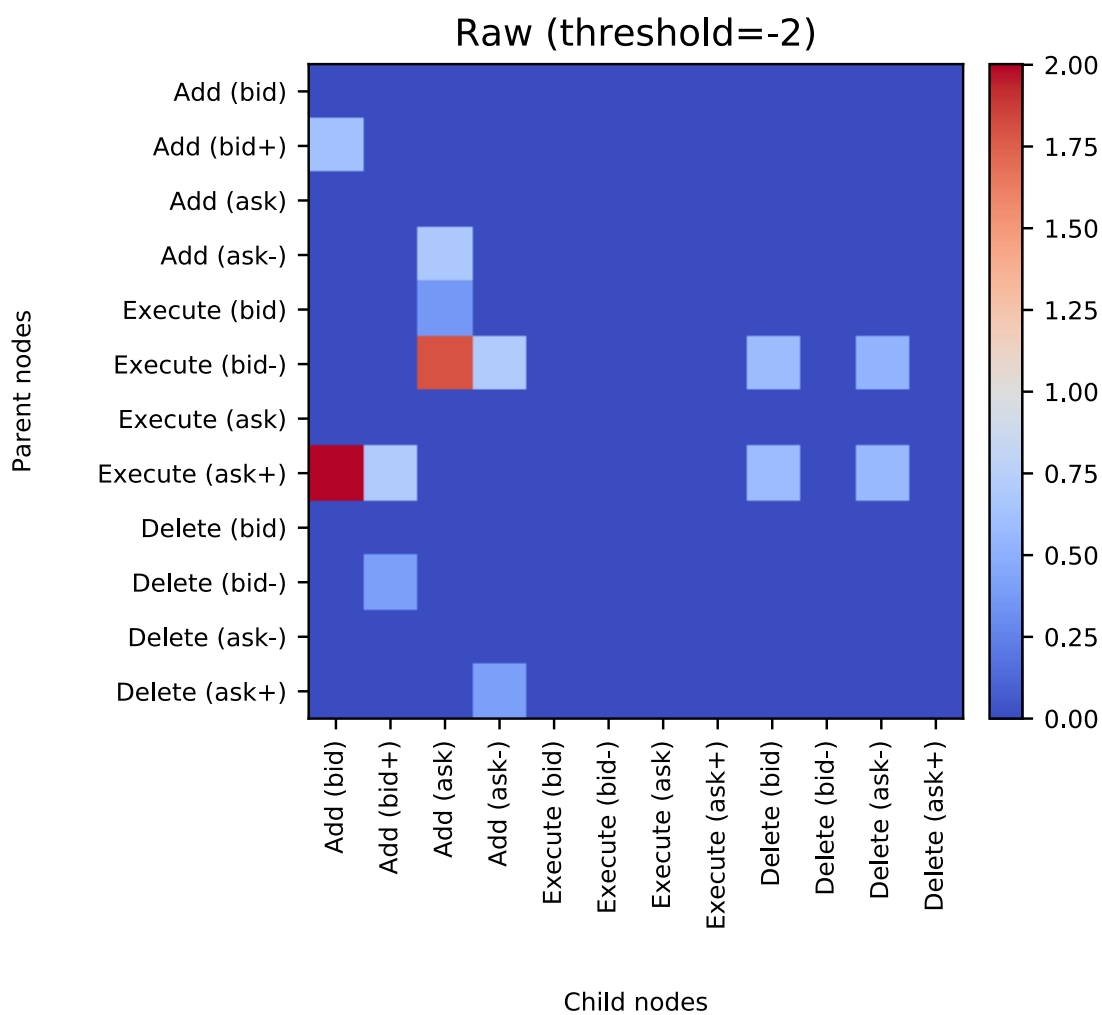


Figure 3.12. Median connection weight point estimates with threshold. The figure shows the median estimate of connections weights after applying a 20% hard threshold.

## 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.
- Hélène 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 predictive 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 limit 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.
- Fabien Guilbaud and Huyền Pham. Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94, 2013.
- 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.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dhharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529 EP –, 02 2015. URL <http://dx.doi.org/10.1038/nature14236>.
- 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.
- David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354 EP –, 10 2017. URL <http://dx.doi.org/10.1038/nature24270>.
- 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.