An Order Flow Model and a Liquidity Measure of Financial Markets

by

Adlar Jeewook Kim

B.S. Carnegie Mellon University (1998)
B.S. Carnegie Mellon University (1998)
M.S. Carnegie Mellon University (2000)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2008

Author ....
Department of Electrical Engineering and Computer Science
August 8, 2008

Certified by ....
Tomaso Poggio
Eugene McDermott Professor
Thesis Supervisor

Certified by ....
Andrew W. Lo
Harris and Harris Group Professor
Thesis Supervisor

Certified by ....
J. Doyne Farmer
Professor, Santa Fe Institute
Thesis Supervisor

Accepted by ....
Terry P. Orlando
Chair, Department Committee on Graduate Students

An Order Flow Model and a Liquidity Measure of Financial Markets

by

Adlar Jeewook Kim

Submitted to the Department of Electrical Engineering and Computer Science
on August 8, 2008, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The thesis seeks a better understanding of liquidity generation process of financial markets and attempts to find a quantitative measure of market liquidity. Various statistical modeling techniques are introduced to model order flow generation, which is a liquidity generation process of the market. The order flow model successively replicates various statistical properties of price returns including fat-tailed distribution of returns, no autocorrelation of returns and strong positive autocorrelation of transaction signs. While attempting to explain how the order flow model satisfies Efficient Market Hypothesis (EMH), I discovered a method of calibrating market liquidity from order flow data.

Thesis Supervisor: Tomaso Poggio
Title: Eugene McDermott Professor

Thesis Supervisor: Andrew W. Lo
Title: Harris and Harris Group Professor

Thesis Supervisor: J. Doyne Farmer
Title: Professor, Santa Fe Institute

To my parents; Jae Won Kim and Jee Sook Hwang

Acknowledgments

My plan to pursue PhD in this topic began many years back when I was an undergraduate student at Carnegie Mellon University. I was one of the fortunate people who had chance to personally interact with the late Herbert Simon who unfortunately is no longer with us today. I would like to thank him for encouraging me to pursue research of my interests even though it could have been (and had been) a difficult journey.

I wish to express my sincere gratitude to all those who helped me during the course of this work. I gratefully acknowledge my mentors Tomaso Poggio, Andrew W. Lo, and J. Doyne Farmer, and my thesis committee members John N. Tsitsiklis and Patrick H. Winston for their patience, guidance and advice.

I am also thankful to all the people at CBCL, LFE, and Santa Fe Institute. At CBCL, I was fortunate to meet people who are now my good friends. I especially thank Nicholas Tung Chan for getting me involved in various projects, Sanmay Das for stimulating discussions which helped me forming ideas for my thesis, and Gadi Geiger for being always available to talk openly about everything over the countless coffee breaks. I was fortunate to share my office with Tony Ezzat who had been a great friend and a colleague. I very much enjoyed talking to Tony about my work and also many other things in life. I also gratefully acknowledge Mary Pat Fitzgerald, Minjoon Kouh, Luis Perez-Breva, Thomas Serre, Christian Shelton, Kathleen Sullivan, and Gene Yeo at CBCL; Ely Dahan at UCLA; Mike Epstein and Svetlana Sussman at LFE; Marilyn Pierce and Janet Fischer at MIT EECS department; Austin Gerig, Nate Metheny, Yonathan Schwarzkopf, Michael Twardos, Javier Vicente, Neda Zamani at Santa Fe Institute for the opportunities to interact with them.

More personally, I am thankful to friends outside of my lab. Especially to Insuk Choi, Eunjong Hong, and Chanryeol Rhyou, your friendships have been invaluable to me. I have to say, we have some good memories while we were at MIT. To my lifelong friends, Hyung Kyu Lee and Hongchul Byun, thank you for keeping me up-to-date on what lives outside of MIT are like.

Finally, to my family. To my brother, Warren Kim, thank you for listening to all my agonies in life and suggesting solutions. I am so lucky to have you as my big brother. To my parents, I thank you for your love and support throughout this journey. None of this could have been possible without you and for that I am dedicating this to you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Modern financial markets provide a vast amount of microscopic data on the behaviors of agents who make buying and selling decisions of financial instruments. For the past few decades, availability of data made it possible to either validate or challenge theoretical models in finance theory through empirical analyses. In this thesis, I propose an extra step towards the better understanding of market microstructure. This is achieved by introducing statistical models that can be trained using the empirical data. Then the resulting models can simulate infinite number of possible scenarios of outcomes, which follow statistical regularities. These models are more realistic than the theoretical ones with fewer realistic assumptions, and can be better tested.

1.1 Major Contributions

This thesis focuses on the modeling of a liquidity generation process of financial markets and derives a quantitative measure for market liquidity. In the first part of this thesis, I demonstrate statistical modeling techniques to model order flow generation, which is a liquidity generation process for financial markets. I implement the model and through simulation, I demonstrate the order flow model successfully replicates various statistical properties of price returns, such as clustered volatility, fat-tailed return distribution, and no predictability of future returns. In the second half of this thesis, I argue that the change in market liquidity explains how my order flow model satisfies the weak form of the Efficient Market Hypothesis (EMH), which asserts that future price returns cannot be predicted from past returns. This is not an obvious result since the order flow model has a predictable transaction sign process (i.e. buyer or seller initiated trades) as one of its components. A method of quantifying market liquidity from order flow data is introduced to explain this result. I also present the periodicity of change in liquidity discovered in the real market data.

1.2 Outline

The following is the outline of my thesis. In Chapter 2, I describe the data set from the London Stock Exchange including the market mechanisms and detailed summary statistics. In Chapter 3 I build a market model based on real market data, which can simulate known statistical properties of the real market. I propose an approach to empirically model the aggregated behavior of the trading crowd, which involves modeling the time sequence of orders placed/canceled in the market called the order flow. In Chapter 4, I propose a parametric model of limit order removal times. The term *removal time* refers to both limit order execution and cancellation times. I show that using a standard statistical technique to model removal times cannot accurately model the removal time's diffusive behavior and therefore introduce a better approach of modeling it. In Chapter 5, I compare the simulated data from our order flow model to the empirical data. For comparison, I focus on three statistical properties of stock returns: fat-tailed distribution of returns, clustered volatility, and no autocorrelation of returns. Although there is no direct link between the statistics of placement/cancellation processes and the statistics of price returns, the results show that the model can successfully replicate the statistical properties of stock returns. In Chapter 6, I introduce a method of calibrating market liquidity from the order flow data. I verify the method by showing the correlation between order sign predictability and the liquidity measure. This explains why the price return series is not autocorrelated when the transact sign series is a long memory process. This result shows how the return series satisfy the efficient market hypothesis (EMH). I also applied the method on real order flow data and identified patterns in change in market liquidity. I conclude in Chapter 7.

Chapter 2

London Stock Exchange Data

All results presented in this thesis are from stocks traded on the London Stock Exchange (LSE) over the period of May 2, 2000 to December 31, 2002. I used LSE data to train/verify the model and to perform various statistical analyses. In this chapter I provide a general overview of the LSE's market mechanism and its data.

Although the market mechanisms differ slightly from one exchange to another, the LSE is one of the largest exchanges in the world; therefore, by studying the LSE we can focus on those characteristics that are common in large exchanges. Also, due to its large data size, we can conduct meaningful statistical analyses. The trades in LSE can occur from two different systems, via electronic limit order book, which is referred to as the on-book market, or via retail service provider (RSP), which is referred to as the off-book quotation market. Empirically it is known that the electronic open limit order book is the dominant mechanism for price formation (LSEbulletin, 2001). The off-book trades mostly consist of large block trades and 75% of them are executed within the best bid/ask prices in the limit order book. In my analysis I used the order, quote, trade data from the Stock Exchange Electronic Trading System (SETS), which is one of the trading services the LSE offers for on-book trading based on the electronic open limit order book. In SETS, upon submission of orders from market participants, the orders are routed to the central order book and are automatically executed. The off-book trades are not used in the analysis because trading times are reported manually, which results in unpredictable reporting time lags. Although the data set only consists of orders that are executed through SETS, it is the dominant mechanism for the price formation, and thus, I believe that it is a sufficient data set for the study.

2.1 Order Matching Mechanism

In the LSE, market participants submit mainly two types of orders to SETS - limit and market. A limit order is an order to transact a prespecified number of shares at a prespecified price. When there is no matching order that can fulfill these conditions, the limit order will remain in the limit order book until one or more matching orders arrive. In contrast, a market order is an order to transact a prespecified number of shares at the best market price[1]. As long as there are matching limit orders, the market orders will be executed immediately as they arrive in the central limit order book. The SETS is an order-driven market, in which there are no market makers[2], and liquidity is solely generated by limit orders submitted by market participants. The participants can choose whether to submit limit or market orders based on their preferences. If they want immediate execution, they will prefer a market order, but there is a risk of having orders executed at undesirable price(s). If they want guaranteed execution price, they will prefer a limit order, but in this case they bear a risk of having to wait a long time until execution, or even of getting no execution at all. When orders arrive in the central limit order book, an execution priority is given to each order. For market orders priority is based on order of arrival. For limit orders priority is based on price and then on order of arrival. The difference between the highest buy and the lowest sell price defines the bid-ask spread. When the size and/or price of an order is modified before execution, a new timestamp is assigned to the order and priority is assigned accordingly. When an order is partially executed, it keeps the original timestamp with no change in execution priority.

2.2 Limit Order Removal Times

As mentioned earlier a market order is guaranteed an immediate execution whereas a limit order is not. While a limit order remains in the order book, it can be either executed or

[1] The best market price refers to the best ask price when it is a buy order and the best bid price when it is a sell order. When the size of a market order is larger than the size of all the limit orders at the best price, the remaining part of the market order, if any, will transact at the next best price.

[2] Market makers are designated person(s) or firm(s) by the exchange to quote bid and ask price in the market.

canceled. In Chapter 3 & 4 we present an order removal times model, which attempts to model limit order execution (time-to-fill) and cancellation (time-to-cancel) times. Both time-to-fill and time-to-cancel describe a time duration of limit order remaining in the order book. Depending on how the order is removed from the book – either by execution or cancellation – the duration is identified accordingly.

2.2.1 Event Time

Throughout my thesis, I will measure removal times in terms of event time. Instead of measuring time in seconds the event time measures it in number of events. An event is defined to be an occurrence that causes change in the order book. In our case, order submissions and cancellations are events, but executions are not. An execution occurs simultaneously with a market order submission and I define such as a single event. This simplifies my studies especially the order flow model in Chapter 3, because I do not have to worry about modeling the actual time intervals between the events. Also other studies have shown that using event time instead of calendar time does not significantly change the results of their analyses (Farmer et al., 2004). Although I do not report in Chapter 4, I have similar results when I model removal times in terms of calendar time.

2.2.2 Time-to-Fill and Time-to-Cancel

When the limit order's price and/or size gets modified before execution, they receive new timestamps. To handle the modification in the most natural way, I simply divide a modified order into multiple orders separated by the time when the limit price and/or size of the order is modified. This is the same as concurrently receiving an order cancellation and a new order submission at the time of the modification. Also, there is no guarantee that limit orders are fully executed at a single point in time. When a limit order is partially executed it becomes ambiguous to define the time-to-fill of that order. To remedy this, we need to consider all the possible scenarios of limit order executions and cancellations. There are four possibilities for limit order execution and cancellation: (1) full execution without partial fill, (2) full execution with partial fill(s), (3) cancellation after partial fill(s), and (4) cancellation without partial fill. Full execution without partial fill and cancellation without partial fill are the simplest cases – they will just be considered as single limit orders with apparent durations for time-to-fill or time-to-cancel. For full execution with partial fill(s), I define the time-to-fill to be the duration for the order to be fully executed regardless of how many partial fills occurred before the full execution. Cancellation with partial fill(s) is more complicated because part of the order generates a time-to-fill and the other part generates a time-to-cancel. In order to resolve this ambiguity, I separate the orders so that one will be a full execution with partial fill(s) and the other will be a cancellation without partial fill. (Both orders are placed at the same original time.) For example, consider a buy order of 2,000 shares at price £10.00 is placed at time $t = 0$, 1,000 shares are executed at $t = 10$, and the rest are canceled at $t = 15$. I treat this as two 1,000 buy orders at price £10.00, placed at $t = 0$. One of these is fully executed at $t = 10$, and the other is canceled at $t = 15$. This separation into multiple orders does not affect price formation.

2.3 Trading Periods: Auction vs. Continuous

The trading day in the LSE contains 3 main trading periods: opening auction, continuous trading, and closing auction[3]. I define auction orders and continuous orders to be orders placed during auction and continuous periods respectively. The market opens with the opening auction at 7:50 am, which ends roughly at 8:00 am. After the opening auction the continuous trading period begins. Then at 4:30 pm the continuous trading period closes and the closing auction begins and lasts for roughly 5 minutes. Unlike continuous orders, auction orders will stay in the order book without execution until the end of the auction period. At the end of each auction period, the LSE calculates a common clearing price using a matching algorithm and executes the matching orders in the limit order book at a common price. In order to prevent traders from manipulating the clearing price of the stock, the auction ending time varies randomly within 30 seconds. In this framework the auction orders' times-to-fill or times-to-cancel are ill defined. When the auction orders are executed at the end of the auction period, they are executed all at once regardless of their placement time. Also, when they are cleared either by execution or cancellation during the continuous trading period (which is followed by the end of the opening auction), the time duration of orders left in the order book during the auction period cannot be included in the calculation of time-to-fill or time-to-cancel since the order matching did not occur continuously during that period. For this reason including auction orders will bias the

[3] The closing auction period was introduced on May 20, 2000.

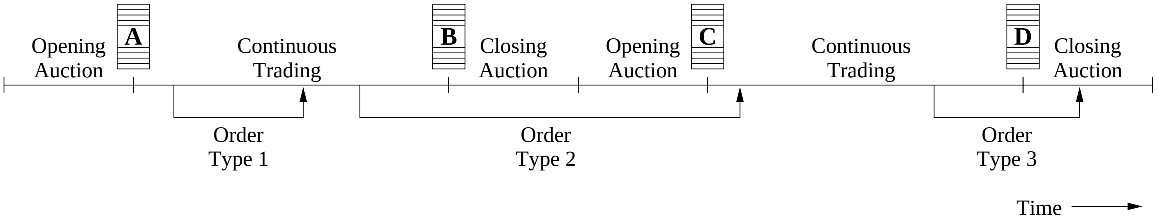


Figure 2-1: Three types of continuous order execution (cancellation) scenarios. The A, B, C, and D indicate different states of limit order book. The arrow for each order indicates order placement time and execution (cancellation) time.

analysis and thus I discard them.

Similarly, limit order removal times can be ill defined for some of the continuous orders as well. Figure 2-1 shows three types of continuous order removal scenarios. Order type 1 is the case where the order is placed during continuous trading and removed in the same period. This is the most common case for limit order removal, and the one in which time-to-fill and time-to-cancel are clearly defined. Order types 2 and 3 correspond to orders placed during the continuous period and carried over to the next trading period. Type 2 corresponds to the case in which an order is placed in one continuous trading period and removed in a subsequent period. One may argue that we can define removing time to be the time that the order remained in the book only during continuous trading. However, this is not an accurate measure since the state of the limit order book at the opening of the closing auction is very different from its state at the closing of the opening auction – indicated by B and C in the figure. In the data set an average of 98% of the orders in the order book during the closing auction are either canceled or executed during the auction. Thus, the states of limit order books B and C in the figure are very different. Lastly, order type 3 indicates carried over continuous orders that are removed during an auction. Similar to auction orders, they are removed during the auction when continuous matching is suspended, so the removal time is ill defined. Among these three types of continuous orders, I only used order type 1 in my analyses and discarded type 2 and 3. In the data set approximately 97% of the continuous orders are type 1 orders.

2.4 Summary Statistics

Table 2.1 & 2.2 show the summary statistics of 9 stocks that I analyzed in the data set. The list consists of stocks with a range of different average trading volumes. Table 2.1 shows that on average 0.21% of orders are auction orders, and 1.87% are type 2 & 3 continuous orders. This shows that discarding auction orders and carried-over continuous orders will have little effect in my analyses. Table 2.1 also shows that type 1 continuous orders are almost equally divided into buy and sell orders. Table 2.2 shows summary statistics of type 1 limit and market orders traded during the continuous trading period.

I categorized limit and market orders in terms of their net effects on the limit order book. The effective market orders are those that resulted in immediate execution. This includes the market orders that transacted a prespecified number of shares at a market price as defined in Section 2.1. Similar to market orders, crossed limit orders are limit orders with a prespecified price that is better than the current opposite best, resulting in immediate execution. Their net effect is the same as that of market orders, and thus we categorize them as effective market orders. Effective limit orders are those that do not result in immediate execution. Effective limit orders consist of canceled limit orders, which are removed from the order book before execution, and filled limit orders, which are executed.

Symbol	# Auction Orders (%)	# Carried Over Orders - Type 2 & 3 (%)	Continuous Orders - Type 1		
			# obs	% buy	% sell
AZN	3,423 (0.14)	28,483 (1.15)	2,444,477	50.52	49.48
BLT	2,977 (0.28)	15,984 (1.51)	1,039,153	51.68	48.32
BSY	3,130 (0.22)	28,682 (2.02)	1,390,002	50.46	49.54
LLOY	3,054 (0.14)	34,405 (1.63)	2,077,984	51.06	48.94
PRU	2,771 (0.18)	23,588 (1.55)	1,496,505	50.39	49.61
RTO	2,984 (0.40)	16,317 (2.16)	738,189	49.77	50.23
RTR	3,141 (0.20)	29,378 (1.89)	1,526,973	50.52	49.48
TSCO	2,885 (0.23)	24,426 (1.91)	1,254,989	50.76	49.24
VOD	3,772 (0.12)	93,909 (3.01)	3,030,508	51.21	48.79

Table 2.1: Summary statistics of auction, carried-over, and continuous orders from May 2, 2000 to December 31, 2002 of 9 individual stocks listed in the LSE. Auction orders are those placed before 8:00 am or after 4:30 pm. The carried-over orders are those placed between 8:00 am and 4:30 pm and not executed or canceled on the same day. Continuous orders are those placed between 8:00 am and 4:30 pm and are either executed or canceled on the same day.

Symbol	Effective Limit Orders			Effective Market Orders		
	# obs	% filled	% canceled	# obs	% market	% crossed
AZN	1,846,940	0.31	0.69	597,537	0.26	0.74
BLT	763,987	0.32	0.68	275,166	0.30	0.70
BSY	1,015,668	0.35	0.65	374,334	0.25	0.75
LLOY	1,451,678	0.39	0.61	626,306	0.32	0.68
PRU	1,086,915	0.34	0.66	409,590	0.32	0.68
RTO	518,675	0.36	0.64	219,514	0.29	0.71
RTR	1,096,126	0.35	0.65	430,847	0.29	0.71
TSCO	847,986	0.44	0.56	407,003	0.31	0.69
VOD	1,958,101	0.50	0.50	1,072,407	0.28	0.72

Table 2.2: Summary statistics of effective limit and effective market continuous orders (Type 1) from May 2, 2000 to December 31, 2002 of 9 individual stocks listed in the LSE. The filled and canceled orders are continuous limit orders that are either executed or canceled. The crossed limit orders are buy(sell) orders with limit price higher(lower) than the ask(bid) price, which result in immediate execution.

Chapter 3

Stock Order Flow Model

3.1 Introduction

In this chapter I introduce a stock order flow model based on the statistical description of order placement and cancellation. Such model is often referred to as “zero intelligence” model since it randomly generates orders without any notion of rationality of agents. This modeling approach contrasts with earlier market models based on the assumption that traders make rational probabilistic decisions based on available information (Glosten and Milgrom, 1985; Kyle, 1985). Surprisingly absence of rationality assumptions in “zero intelligence” models have shown to be capable of replicating various market behaviors.

3.1.1 Related Work

The order flow model introduced in this chapter is much in the same spirit as in a long list of other market models, which attempt to describe order flow as a statistical process. Similar to the model based on rationality assumption, they show a “zero intelligence” model can be used to make strong predictions based on a compact set of assumptions. For instance, Mendelson (1982) studied the behavior of prices and quantities from a model with random order placement with periodic clearing suggesting the possible implications of market conditions. Maslov (2000) studied a simple model of a limit order-driven market, which traders randomly choose to place different types of orders. He showed that despite of such minimalistic settings, the price pattern generated by the model has realistic features such as power law tails of the returns and volatility clustering. Challet and Stinchcombe (2001) introduced a model of limit order markets based on the Island ECN order book data. From the model they observed power law tails of the returns and volatility clustering. Bouchaud et al. (2002) studied statistical properties of the order book of stocks traded in the Paris Bourse. They found the statistics of limit order prices follow power-law around the current bid-ask quotes, and the shape of average order book can be reproduced using simple "zero intelligence" model. Smith et al. (2003) and Daniels et al. (2003) introduced a model which the order arrival and cancellation are defined to be Poisson random processes. They found that such simplified market model is capable of inducing anomalous diffusion and temporal structure of prices. Later Farmer and Mike (2006) further improved this model with different order placement and cancellation processes and show that the model is capable of replicating statistical properties such as distribution of the volatility and the bid-ask spread of some stocks they used for training.

3.1.2 Our Approach

I introduce a new approach of modeling order flow, which is complementary to the models studied earlier. In the earlier models, various unrealistic assumptions were made such as neglecting long memory of order signs, or overlooking complexity of order placement and cancellation processes. Although some simplification is inevitable, I attempt to create an order flow model closer to reality by applying various statistical modeling techniques.

The order flow model represents placement or cancellation of four types of orders - market[1] buy, market sell, limit[2] buy, and limit sell orders. There are three variables, which are used to model such orders:

1. Order sign ψ: A variable indicating whether the order is buy or sell order.

2. Order price p: A variable for the price of an order. An order is considered a market order if the price generated by the model crosses the opposite best price[3] in the order book.

3. Cancellation time τ: A variable indicating how long the order will remain in the order book for execution. If an order is not executed within this time period, it is canceled. The event time is used and this only applies to limit orders since all market orders

[1] A market order is an order to buy or sell a stock at the current market price.

[2] A limit order is an order to buy or sell a security at a specific price.

[3] A best bid price for a sell order generated and a best ask price for a buy order generated.

gets executed immediately except for unrealistic market state where no limit orders are present in the order book.

I made couple of simplifications in my order flow model. First, the model only simulates orders consist of all three components listed above. Some of the simulated orders will have price crossing the opposite best price, which are identified as effective market orders resulting immediate executions. These orders are similar to ordinary market orders with no price embedded in their order description. Therefore when I train the price model, I cannot ignore market orders with missing price. Instead, I attached hypothetical prices to market orders based on the price at which the market order is executed. I will go over this in more detail when I describe the price model in Section 3.3. Second, an order size is fixed in the model. This is justified by the study of the on-book market of the LSE by Farmer et al. (2004). The study showed that orders that remove more than the depth at the opposite best quote are rare. Similarly, Bouchaud et al. (2002) also showed in his "zero intelligence" model that the size distribution play a minor role in reproducing the shape of average order book. Third, the model simulates order flow in event time. The event in event time is defined to be either arrival or cancellation of an order. This implies the model does not require to simulate order arrival times. In reality the actual calendar time (such as in seconds) between the events will not be uniform, however, simulating order flows in event time will not change the price formation.

In the rest of this chapter I describe each component – order sign, price, and cancellation time – of the model in more detail.

3.2 Order Sign Model

In the previous studies by Lillo and Farmer (2004) and Bouchaud et al. (2004), the long memory effect of supply and demand in the market data is introduced. In the study, they analyzed data from the London Stock Exchange, the Paris Bourse, and the New York Stock Exchange and showed that the strong autocorrelation of buyer/seller initiated trades[4] are universal. This was due to a stream of buy and sell orders being autocorrelated, which implies a buy order is likely to arrive in the market after stream of buy orders and vice

[4] Buyer initiated trades are trades occurred by market buy orders and seller initiated trades are trades occurred by market sell orders.

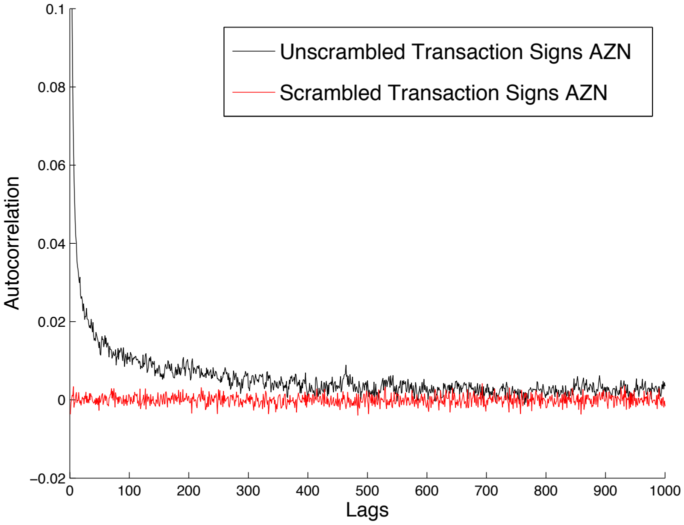


Figure 3-1: Autocorrelation of transaction sign series for the stock AZN (black). The same transaction sign series is randomly reordered to kill the autocorrelation (red).

versa. Figure 3-1 shows the autocorrelation function of transaction sign series for the stock AZN in black (+1 for buyer and −1 for seller initiated orders). If the model assumes IID order signs, the autocorrelation function of simulated transaction signs will look similar to the plot shown in red in Figure 3-1 where the original transaction signs are randomly reordered.

The Hurst exponent of the transaction sign series for AZN is 0.6753, which indicates the long memory process[5]. In the following section we introduce fractional autoregressive moving average model (FARIMA), which can simulate long memory process of transaction sign series.

3.2.1 FARIMA Model

In order to model strongly autocorrelated stream of buy and sell initiated trades, we used fractional autoregressive integrated moving average (FARIMA) model introduced by Granger and Joyeux (1980) and Hosking (1981). FARIMA (sometimes called fractional ARIMA or ARFIMA) is a generalization of autoregressive moving average model (ARIMA) introduced by Box and Jenkins (1970), which is generally referred to as an ARIMA(p, d, q) with integer values $p, d, q \geq 0$ where p, d, q are the order of the autoregressive, integrated,

[5] If the Hurst exponent is $0.5 < H < 1$, the process is a long memory process.

and moving average part respectively. Given the time series data Ψ_t where Ψ_t's are real numbers at time t, ARIMA(p, d, q) model is given by

$$\left(1 - \sum_{i=1}^{p} \phi_i L^i\right) (1 - L)^d \Psi_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \epsilon_t, \tag{3.1}$$

where L is the lag operator, ϕ_i and θ_i are the parameters of the autoregressive and moving average part of the model respectively, and ϵ_t are the error terms with zero mean and variance σ_ϵ^2.

In Equation 3.1 the level of term-by-term differencing is expressed with the lag operator L and integrated parameter d. Since d is defined to be an integer greater than or equal to 0, $(1 - L)^d$ can be expressed as

$$(1 - L)^d = \sum_{k=0}^{d} \binom{d}{k} (-1)^k L^k, \tag{3.2}$$

where the binomial coefficients can be expressed in terms of gamma function $\Gamma(\cdot)$ as

$$\binom{d}{k} = \frac{d!}{k!(d-k)!} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}. \tag{3.3}$$

FARIMA models are a natural extension of the ARIMA models by allowing the parameter d of integrated part of the model to be a real value within the range of $-\frac{1}{2} < d < \frac{1}{2}$. Since $\Gamma(\cdot)$ is defined for all real numbers, Equation 3.3 can be used to extended the model for the real numbers of this range. As shown by Granger and Joyeux (1980) and Hosking (1981), the range of d that is interesting in the context of long memory process is $0 \leq d < \frac{1}{2}$. Also parameters p and q allow for more flexible modeling of short range properties whereas the parameter d determines the long term behavior.

In my model, I use FARIMA$(0, d, 0)$, which only focuses on the long term behavior of the time series data. When we set $p, q = 0$, the FARIMA$(0, d, 0)$ process can be represented as an infinite moving average or an infinite autoregressive process with coefficient that can be given explicitly (Hosking, 1981). This result becomes useful when we want to predict future observations from a close to infinite past. For the prediction from a finite past of size

k (which can be a very large number), the best linear predictor $\hat{\Psi}_t$ is

$$\hat{\Psi}_t = \sum_{j=1}^{k} \beta_{kj} \Psi_{t-j}, \tag{3.4}$$

where β_{kj} is

$$\beta_{kj} = -\binom{k}{j} \frac{\Gamma(j-d)\Gamma(k-d-j+1)}{\Gamma(-d)\Gamma(k-d+1)}. \tag{3.5}$$

Also the estimated parameter d of FARIMA process has a following relation

$$\hat{d} = \hat{H} - \frac{1}{2}, \tag{3.6}$$

where $\hat{H}$ is estimated Hurst exponent of the time series data. This can be shown by deriving spectral density of FARIMA model as a function of spectral density of ARMA model which is equal to FARIMA$(p, 0, q)$. The detailed derivation of this can be found in Hosking (1981) and Beran (1994, Chapter 2).

For my order sign model, I first model transaction sign series of a stock – series of buyer/seller initiated trades – with FARIMA$(0, d, 0)$. In this case, time series variable Ψ_t holds $+1$ (-1) for buyer (seller) initiated transactions. The parameter $\hat{d}$ of FARIMA model is derived from Equation 3.6 using the estimated Hurst exponent $\hat{H}$ of transaction sign series data. In the rest of this section I explain how the order signs can be generated from this model.

3.2.2 Order Sign Generation

From a FARIMA$(0, \hat{d}, 0)$ model trained with transaction signs, I can simulate an order sign at time t using transaction sign predictor $\hat{\Psi}_t$. This can be computed from Equation 3.3 using previously generated transaction signs. If we represent a buyer and a seller initiated order as $+1$ and -1 respectively, the transaction sign predictor $\hat{\Psi}_t$ is

$$\hat{\Psi}_t = (+1)P(buy) + (-1)P(sell), \tag{3.7}$$

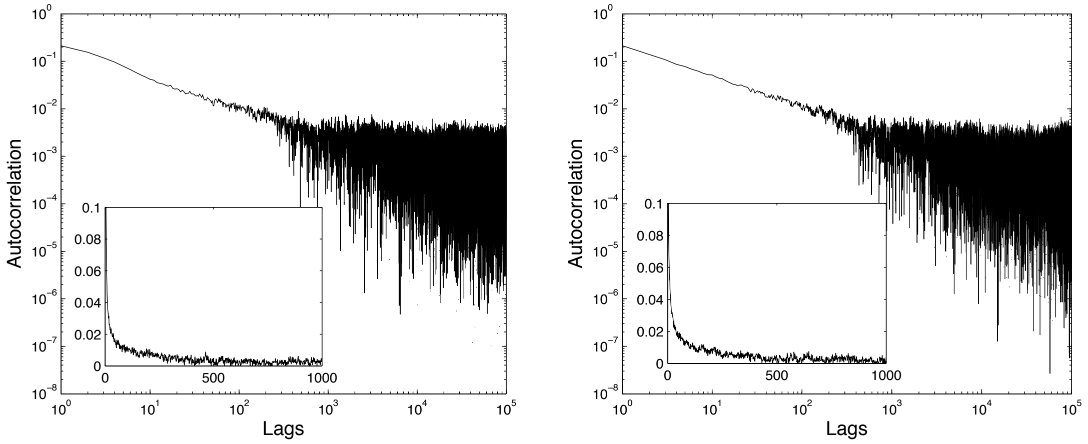


Figure 3-2: Autocorrelation of the actual (left) and the simulated (right) transaction sign series for stock AZN. The FARIMA$(0, \hat{d}, 0)$ model with $\hat{d} = 0.1753$ is used. 10,000 past order signs are used to compute the sign predictor $\hat{\Psi}_t$ using Equation 3.4. 568,563 transaction signs are simulated from the model which is equal to the actual number of transaction signs in the data set.

and the probability of transaction being a buyer initiated at time t is

$$P(buy) = \frac{\hat{\Psi}_t + 1}{2}. \tag{3.8}$$

Then I generate order sign ψ_t (buy or sell) based on this probability. Note that transaction (buyer or seller initiated) will only occur when generated order is an effective market order. If effective limit order is generated, the transaction series will not change, therefore, the probability of order sign will not change. This way, the simulation will generate long memory transaction sign series similar to the stock data used for training.

Figure 3-2 shows autocorrelation functions of actual and simulated transaction signs for the stock AZN in both linear and log-log scale. The FARIMA$(0, \hat{d}, 0)$ with $\hat{d} = 0.1753$ is used to compute order sign probabilities. To calculate the transaction sign predictor $\hat{\Psi}_t$, past 10,000 signs are used. The figure shows the FARIMA model successfully generates transaction signs with long memory similar to the actual data. The estimated parameter of order sign model for all 9 stocks are shown in Appendix A.1. It shows the parameter estimates $\hat{d}$ for remaining 8 stocks are also within the range of $0 \leq d < \frac{1}{2}$ implying long memory processes.

3.3 Price Model

In this section I introduce an architecture of modeling order price generation using a variant of Hidden Markov Models (HMM) called an Input-Output Hidden Markov Model (IOHMM) (Rabiner, 1989; Bengio, 1999). IOHMMs are well suited model to generate prices sequentially over time from conditional distributions. Before I go into details about the IOHMMs I will address the statistical properties of time series data of order prices observed from the data.

The order price is defined as a difference between logarithm of order price and the logarithm of same best price. Thus, the order price p_t for a limit order placed at time t is defined as

$$p_t = \begin{cases} \log(p_{bid,t}) - \log(p_{l,t}) & \text{if buy order} \\ \log(p_{l,t}) - \log(p_{ask,t}) & \text{if sell order.} \end{cases} \tag{3.9}$$

For example when a limit buy order of £5.10 is placed in the market and the current bid price in the market is £5.11, then the price is the difference of the logarithm of current bid price and the logarithm of its limit price, which is 0.002 in this case. Similarly the order price for limit sell order is calculated by taking a difference of the logarithm of limit price and the logarithm of current ask price. By this definition having negative (positive) p_t refers to more (less) aggressive orders relative to the same best price.

In the data set, not all orders have price information. Market order is an order without a prespecified price which gets executed immediately. For these orders I attached a hypothetical price by referring to price at which the orders are executed. When a market order is executed at more than one price by removing more than the depth at the opposite best price, I define hypothetical price to be the latest executed price of that order. For example when a part of market buy order gets executed at ask price of £5.10 and rest at next ask price of £5.11, the hypothetical price for this order is £5.11. The rationale behind this is such market order has the same effect as limit order with the same order size and the hypothetical price as its limit price.

Figure 3-3 shows the empirical probability density functions (PDFs) of order prices of stock AZN in semi-log scale. The figure shows the price distributions for buy and sell orders are indifferent. Also the distributions are asymmetric indicating given the same magnitude

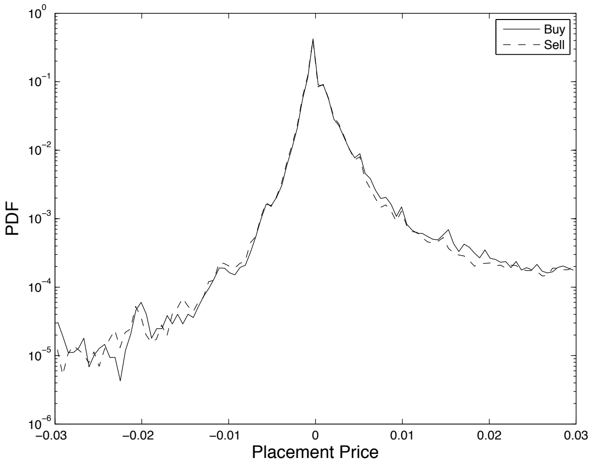


Figure 3-3: Empirical PDFs of order price for buy (line) and sell (dotted line) orders of stock AZN.

of price $|p_t|$, relatively smaller number of aggressive orders ($p_t < 0$) are placed compare to the non-aggressive orders ($p_t > 0$).

In addition to asymmetric distribution, the data show that bid-ask spread is relevant to price distributions. The bid-ask spread at time t is defined as a difference between the logarithm of ask and the logarithm of bid prices at time t,

$$x_t = \log(p_{ask,t}) - \log(p_{bid,t}). \tag{3.10}$$

Figure 3-4 shows the empirical PDFs of stock AZN conditioning on bid-ask spread x_t for buy (left) and sell (right) orders. For both buy and sell, the PDFs for orders placed while the spread is tight ($x_t \leq 0.0025$) has a heavier tail than the PDFs for order placed while the spread is wide ($x_t > 0.0025$). This implies order price p_t is dependent of spread x_t. Also this has an interesting implication regarding traders' order placing behavior. When bid-ask spread is wide, it shows that traders seeking immediacy will place aggressive orders with higher probability than the time when bid-ask spread is tight. Also for traders seeking better price will place less aggressive orders with higher probability when the bid-ask spread is wide.

The order price model is to simulate price of orders forming asymmetric distribution as shown in Figure 3-3, which also has dependence to spread as shown in Figure 3-4. Since I am modeling sequence of order prices dependent on bid-ask spread, I use a variant of

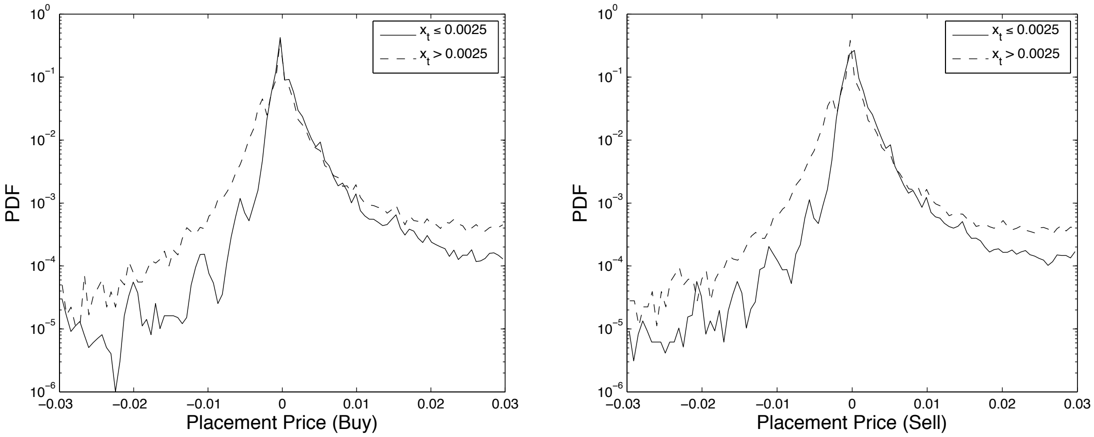


Figure 3-4: Empirical PDFs of order price of stock AZN conditioning on bid/ask spread at time t defined as x_t. The PDFs for orders placed while $x_t \leq 0.0025$ and $x_t > 0.0025$ are plotted separately for buy orders (left) and sell orders (right).

Hidden Markov Model called Input-Output Hidden Markov Model (IOHMM).

3.3.1 Input-Output Hidden Markov Models (IOHMMs)

A Hidden Markov Model (HMM) is a probabilistic model widely used to learn data with sequential structure. The sequential structure exists in various applications and HMMs have been applied to problems such as natural language processing, handwriting recognition, pattern recognition in molecular biology, fault detection, and gesture recognition to name the few.

The HMM assumes the system being modeled to be a Markov process with unknown number of "hidden" states. In order to model order prices sequentially placed in the market that is also dependent on bid-ask spread, I use a variant of HMM called an Input-Output Hidden Markov Model (IOHMM). The IOHMM presents similarities to HMMs, but it also allows to map input sequences to output sequences using the same processing style as recurrent neural networks. Unlike recurrent neural networks, IOHMMs consider a probability distribution over a *discrete* state dynamical system instead of continuous state. The IOHMM is defined by

- m: A number of discrete states.

- $P(s_0)$: Initial state distribution where state at time 0 to be $s_0 = s_1, ..., s_m$.

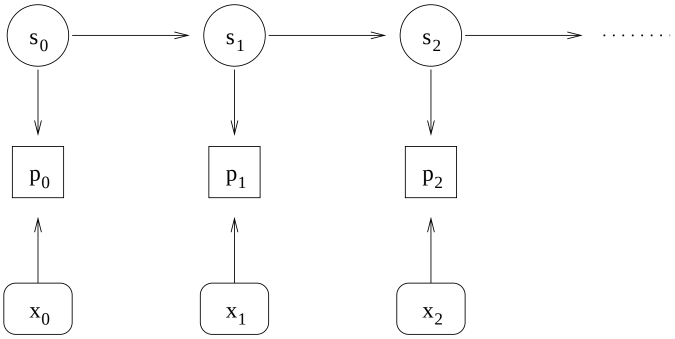


Figure 3-5: Bayesian network describing IOHMM for order placement price model. s_t, p_t, and x_t in the diagram refer to hidden state, price, and bid-ask spread at time t respectively.

- $P(s_{t+1}|s_t)$: State transition probabilities from s_t to s_{t+1}.
- $P(p_t|s_t, x_t)$: Output probability of output p_t given state s_t and input value x_t at time t. The state variable s_t is discrete and the input variable x_t can be both discrete and continuous. In the price model, I use conditional mixture Gaussian for the output probability with continuous conditional variable x_t, which is a bid-ask spread at time t.

In addition to state and output variables s_t and p_t, IOHMM has additional input variable x_t denoting an input value at time t. Similar to HMMs, IOHMMs are simplified by introducing two crucial conditional independence assumptions.

$$\begin{aligned} P(s_{t+1}|s_0, s_1, ..., s_t, p_0, p_1, ..., p_t, x_0, x_1, ..., x_t) &= P(s_{t+1}|s_t), \\ P(p_t|s_0, s_1, ..., s_t, p_0, p_1, ..., p_{t-1}, x_0, x_1, ..., x_t) &= P(p_t|s_t, x_t), \end{aligned} \tag{3.11}$$

which implies the state variable s_t will summarize all the past values of the state sequences and both state and input variables s_t and x_t will summarize all the past values of the output sequences.

Figure 3-5 shows the Bayesian network describing IOHMM for the order price model. A variable s_t indicating the state at time t can be loosely defined as a market state and with additional market information of bid-ask spread x_t, the model outputs price p_t. The figure describes the conditional independence assumptions described in Equation 3.11 where given a state and a bid-ask spread at time t (indicated with variables s_t and x_t respectively), the probability of price p_t is independent from previous market states and bid-ask spreads.

Similarly, the model describes only current state s_t is relevant to determine the next state s_{t+1}, in other words, the transition of next state is independent from all the other previous bid-ask spreads and states given the current market state.

3.3.2 A Supervised Learning Algorithm

Training of IOHMM price model is similar to ordinary HMM model except the output probability is a conditional distribution over not only the given state s_t but also the real-valued input variable of bid-ask spread x_t. Like HMM, learning of IOHMM is derived from the maximum likelihood principle. Given the training data of N pairs of input-output sequences, which in our case are the prices p_t's and bid-ask spreads x_t's of a particular stock, the likelihood function is given by

$$L(\theta) = \prod_{n=1}^{N} P(p_1^{T_n} | x_1^{T_n}, \theta), \tag{3.12}$$

where θ denotes set of parameters for initial state $P_0(s_0)$, state transition $P_1(s_{t+1}|s_t)$, and output $P(p_t|s_t, x_t)$ probability distributions of IOHMM. In the training set, each input-output sequence represents the price $p_1^{T_n}$ and bid-ask spread $x_1^{T_n}$ within a single trading day. The data set has 675 days of trading data ($N = 675$) and each day has variable number of input-output pairs – for example, a stock AZN has T_n ranging from 446 to 9,682 pairs.

The maximization of Equation 3.12 involves parameter estimation with missing data, where the missing (or hidden) variable is the state sequence $s_1, ..., s_{T_n}$ for each input-output sequence $1, .., N$. I use Expectation-Maximization (EM) algorithm to maximize the above likelihood function with hidden states. The EM algorithm formalized in Dempster et al. (1977) solves maximum likelihood estimation (MLE) problem using an iterative approach consisting of two parts: an estimation step (E-step) and a maximization step (M-step). Before breaking down the algorithm into these two steps it is necessary to formulate how EM computes the likelihood of IOHMM of Equation 3.12 when the state sequences are not observable from the training data. The derivation of this problem for more generalized version of IOHMM where the transition probabilities also have dependencies on input values is described in Bengio and Frasconi (1995) and Bengio and Frasconi (1996).

The EM algorithm attempts to estimate complete data likelihood, which consist of input, output, and state sequences. Assuming we have a complete data D_c including the state sequences and set of parameters for IOHMM θ, the complete data likelihood is defined as

$$L_c(\theta; D_c) = \prod_{n=1}^{N} P(p_1^{T_n}, s_1^{T_n} | x_1^{T_n}; \theta). \tag{3.13}$$

Using IOHMM's conditional independence assumption stated in Equation 3.11, we can rewrite Equation 3.13 as following

$$\begin{aligned} L_c(\theta; D_c) &= \prod_{n=1}^{N} \prod_{t=1}^{T_n} P(p_t | s_t, x_t; \theta) P(s_t | s_{t-1}; \theta) \qquad (3.14) \\ &= \prod_{n=1}^{N} \prod_{t=1}^{T_n} \prod_{i=1}^{S} \prod_{j=1}^{S} P(p_t | s_t = i, x_t; \theta)^{z_{i,t}} P(s_t = i | s_{t-1} = j; \theta)^{z_{i,t}, z_{j,t-1}}, \end{aligned}$$

where $z_{i,t}$ is an indicator variable that $z_{i,t} = 1$ if the state at time t is i and $z_{i,t} = 0$ otherwise. Since the value of state sequence is not observable from the training data, I cannot directly compute the complete data likelihood. Instead I formulate *expected* complete data log-likelihood using auxiliary function which integrates the logarithm of likelihood function (Equation 3.14) over the distribution of state paths $\mathcal{S} = \{s_1^{T_n}; n = 1, ..., N\}$ using the parameter estimate $\hat{\theta}$ from the previous iteration[6]. Given the observable output and input sequences $\mathcal{P} = \{p_1^{T_n}; n = 1, ..., N\}$ and $\mathcal{X} = \{x_1^{T_n}; n = 1, ..., N\}$, and previously estimated model parameter $\hat{\theta}$, the auxiliary function $Q(\theta; \hat{\theta})$ is defined as

$$\begin{aligned} Q(\theta; \hat{\theta}) &= E_{\mathcal{S}} \left[l_c(\theta; D_c) | \mathcal{P}, \mathcal{X}, \hat{\theta} \right] \qquad (3.15) \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{i=1}^{S} E_S \left[z_{i,t} | p_1^T, x_1^T, \hat{\theta} \right] \log P(p_t | s_t = i, x_t; \theta) \\ &\quad + \sum_{j=1}^{S} E_{\mathcal{S}} \left[z_{i,t}, z_{j,t-1} | p_1^T, x_1^T, \hat{\theta} \right] \log P(s_t = i | s_{t-1} = j; \theta) \\ &= \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{i=1}^{S} \hat{\gamma}_{i,t} \log P(p_t | s_t = i, x_t; \theta) \\ &\quad + \sum_{j=1}^{S} \hat{\xi}_{ij,t} \log P(s_t = i | s_{t-1} = j; \theta) \end{aligned}$$

[6] For the first iteration $\hat{\theta}$ is picked randomly.

where given the model and the observed input-output sequences, $\hat{\gamma}_{i,t}$ is defined as the probability of being in state i at time t, and $\hat{\xi}_{ij,t}$ is defined as the probability of being in state j at time $t-1$, and state i at time t. The following formulates these two probability measures

$$\begin{aligned} \hat{\gamma}_{i,t} &= P(s_t = i | p_1^T, x_1^T; \hat{\theta}); \\ \hat{\xi}_{ij,t} &= P(s_{t-1} = j, s_t = i | p_1^T, x_1^T; \hat{\theta}). \end{aligned} \tag{3.16}$$

Note that Equation 3.16 can be computed with known input-output sequences p_1^T and x_1^T, and previously estimated parameter $\hat{\theta}$. In order to compute $\hat{\gamma}_{i,t}$ and $\hat{\xi}_{ij,t}$ efficiently, I use forward-backward procedure introduced in Baum and Egon (1967) and Baum and Sell (1968). The forward probability $\alpha_{i,t}$ is defined as

$$\begin{aligned} \alpha_{i,t} &= P(p_1^t, s_t = i | x_1^t) \\ &= \sum_{j=1}^{S} P(p_1^t, s_t = i, s_{t-1} = j | x_1^t) \\ &= \sum_{j=1}^{S} P(p_t | p_1^{t-1}, s_t = i, s_{t-1} = j, x_1^t) \\ &\quad \cdot P(s_t = i | p_1^{t-1}, s_{t-1} = j, x_1^t) P(p_1^{t-1}, s_{t-1} = j | x_1^t) \\ &= \sum_{j=1}^{S} P(p_t | s_t = i, x_t) P(s_t = i | s_{t-1} = j) \alpha_{j,t-1} \end{aligned} \tag{3.17}$$

where the simplified forward equation consists of output and state transition probabilities of the IOHMM. The last step of Equation 3.17 is using conditional independence assumption of the model shown in Equation 3.11 to simplify the computation. It also shows that forward probability $\alpha_{i,t}$ is computed recursively starting with initial forward probability at time 0 equal to the initial state probability: $\alpha_{i,0} = P(s_t = i)$.

Similar to forward probability, a backward probability $\beta_{i,t}$ is defined as

$$\begin{aligned} \beta_{i,t} &= P(p_{t+1}^T | s_t = i, x_t^T) \\ &= \sum_{j=1}^{S} P(p_{t+1}^T, s_{t+1} = j | s_t = i, x_t^T) \\ &= \sum_{j=1}^{S} P(p_{t+1} | p_{t+2}^T, s_{t+1} = j, s_t = i, x_t^T) \\ &\quad \cdot P(p_{t+2}^T | s_{t+1} = j, s_t = i, x_t^T) P(s_{t+1} = j | x_t = i, u_t^T) \\ &= \sum_{j=1}^{S} P(p_{t+1} | s_{t+1} = j, x_t) P(s_{t+1} = j | s_t = i) \beta_{j,t+1} \end{aligned} \tag{3.18}$$

where the backward probability can also be computed recursively starting with initial condition $\beta_{i,T} = 1$. From this we can see that the forward and backward procedures for the IOHMM is similar to HMM, except additional input condition is included.

Using Equations 3.17 and 3.18, I can compute $\gamma_{i,t}$ and $\xi_{ij,t}$ as following using the forward, backward, output and transition probabilities as following

$$\begin{aligned} \gamma_{i,t} &= P(s_t = i | p_1^T, x_1^T) \\ &= P(s_t = i, p_1^T | x_1^T) / P(p_1^T | x_1^T) \\ &= P(p_1^t, s_t = i | x_1^t) P(p_{t+1}^T | s_t = i, x_t^T) / P(p_1^T | x_1^T) \\ &= \frac{\alpha_{i,t} \beta_{i,t}}{\sum_{j=1}^{S} \alpha_{j,t} \beta_{j,t}}, \end{aligned} \tag{3.19}$$

$$\begin{aligned} \xi_{ij,t} &= P(s_{t-1} = j, s_t = i | p_1^T, x_1^T) \\ &= P(s_{t-1} = j, s_t = i, p_1^T | x_1^T) / P(p_1^T | x_1^T) \\ &= P(p_1^{t-1}, s_{t-1} = j | x_1^t) \cdot P(p_t | s_t = i, x_t) \cdot P(s_t = i | s_{t-1} = j) \\ &\quad \cdot P(p_t^T | s_t = i, x_t^T) / P(p_1^T | x_1^T) \\ &= \frac{\alpha_{j,t-1} P(p^t | s_t = i, x_t) P(s_t = i | s_{t-1} = j) \beta_{i,t}}{\sum_{j=1}^{S} \alpha_{j,t} \beta_{j,t}}. \end{aligned} \tag{3.20}$$

As mentioned earlier the EM algorithm is an iterative approach consisting of two parts:

- E-step (Estimation): The algorithm computes posterior probabilities $\hat{\gamma}_{i,t}$ and $\hat{\xi}_{ij,t}$ for all states i and time t using previously estimated model parameter $\hat{\theta}$ and set of input-output training sequences.

- M-step (Maximization): The algorithm searches for new model parameter $\hat{\theta}'$ that maximizes the auxiliary function $Q(\theta; \hat{\theta})$.

The algorithm iterates the expectation and maximization steps until the improvement of expected complete log-likelihoods over iteration falls below certain threshold.

For the M-step, maximization of parameter θ depends on the choice of models for transition and output probabilities. But regardless of the choice of models, the auxiliary function (Equation 3.15) can be broken down into two separate maximization problems. For each state $j = 1, .., S$, I find parameter estimate θ_1 for transition probability where it optimizes the following maximization problem,

$$\max_{\theta_1} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{i=1}^{S} \hat{\xi}_{ij,t} \log P(s_t = i | s_{t-1} = j; \theta_1). \tag{3.21}$$

Similarly for output probabilities, I find parameter θ_2 that optimizes the following maximization problem for each state j

$$\max_{\theta_2} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\gamma}_{j,t} \log P(p_t | s_t = j, x_t; \theta_2). \tag{3.22}$$

In the rest of this section I describe the M-step specific to the choice of models of output and transition probabilities.

State Transition Probability

For my price model, solving Equation 3.21 is simple because the transition probability does not depend on real-valued bid-ask spread x_t. Since the state variable s_t are discrete and finite, the transition probability behaves like lookup tables addressed by the state symbols. I use Lagrange multiplier λ to find transition probability values, which maximizes Equation 3.21 with constraint that transition probability from state j to all i's sums to 1. Let's call $a_{ij} = P(s_t = i | s_{t-1} = j)$, then

$$\frac{\partial}{\partial a_{ij}} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{i=1}^{S} \hat{\xi}_{ij,t} \log a_{ij} + \lambda \left(1 - \sum_{i=1}^{S} a_{ij} \right) \quad = \quad \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\xi}_{ij,t}}{a_{ij}} + \lambda. \tag{3.23}$$

By setting Equation 3.23 equal to 0, we get

$$\hat{a}_{ij} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_{ij,t}}{\lambda} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \xi_{ij,t}}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \gamma_{j,t}} \tag{3.24}$$

Output Probability

Solving Equation 3.22 requires bit more work. For each state in the model, output probability describes conditional distribution of prices conditioning on the market's bid-ask spread. For this I used mixture of conditional Gaussian distributions. Given state i and mixture k, distribution for output probability of the model is

$$P(p_t|s_t = j, m_t = k, x_t) = \frac{1}{\sigma_{jk}\sqrt{2\pi}} \exp\left(-\frac{(p_t - B_{jk}x_t - \mu_{jk})^2}{2\sigma_{jk}^2}\right), \tag{3.25}$$

where parameters μ_{jk}, B_{jk}, and σ_{jk}^2 refers to mean, regression coefficient for conditional variable, and variance of the conditional Gaussian of state j and mixture k. As shown in Equation 3.25, defining output probability as Gaussian mixture requires additional discrete mixture variable m_t. This requires rewriting the posterior probability in Equation 3.22 to incorporate the mixture variable. The new posterior $\omega_{jk,t}$ for state j, mixture k at time t can be derived as

$$\begin{aligned} \omega_{jk,t} &= P(s_t = j, m_t = k|p_1^T, x_1^T) \qquad (3.26) \\ &= \frac{P(p_t|s_t = j, m_t = k, p_1^{t-1}, p_{t+1}^T, x_1^T)P(s_t = j, m_t = k|p_1^{t-1}, p_{t+1}^T, x_1^T)}{p(p_t|p_1^{t-1}, p_{t+1}^T, x_1^T)} \\ &= \frac{P(p_t|s_t = j, m_t = k, x_t)P(m_t = k|s_t = j)P(s_t = j|p_1^{t-1}, p_{t+1}^T, x_1^T)}{P(p_t|p_1^{t-1}, p_{t+1}^T, x_1^T)} \\ &= \gamma_{j,t} \cdot \frac{P(p_t|s_t = j, m_t = k, x_t)P(m_t = k|s_t = j)}{\sum_{l=1}^{K} P(p_t|s_t = j, m_t = l, x_t)}, \end{aligned}$$

which is a function of posterior $\gamma_{j,t}$ defined in Equation 3.19, and posterior mixture and output probabilities with parameters B_{jk}, μ_{jk}, and σ_{jk}^2. This results re-writing our expected complete log-likelihood in Equation 3.22 as

$$l = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{M} \hat{\omega}_{jk,t} \log \frac{1}{\sigma_{jk}\sqrt{2\pi}} - \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{M} \hat{\omega}_{jk,t} \left(\frac{(p_t - B_{jk}x_t - \mu_{jk})^2}{2\sigma_{jk}^2}\right). \tag{3.27}$$

In M-step we compute parameter estimates for output probabilities, which maximize the above log-likelihood. For output probability of state j and mixture k, differentiating the new log-likelihood respect to parameters B_{jk}, μ_{jk}, and σ_{jk}^2 result following equations

$$\begin{aligned} \frac{\partial l}{\partial B_{jk}} &= \frac{1}{\sigma_{jk}^2} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}(p_t - B_{jk}x_t - \mu_{jk})x_t = 0, \\ \frac{\partial l}{\partial \mu_{jk}} &= \frac{1}{\sigma_{jk}^2} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}(p_t - B_{jk}x_t - \mu_{jk}) = 0, \\ \frac{\partial l}{\partial (\sigma_{jk}^2)^{-1}} &= \sigma_{jk}^2 \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t} - \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}(p_t - B_{jk}x_t - \mu_{jk})^2 = 0. \end{aligned} \tag{3.28}$$

For each output probability for state j and mixture k, the set of new parameter estimates can be computed by solving Equation 3.28 with posteriors $\hat{\omega}_{jk,t}$, $\hat{B_{jk}}$, $\hat{\mu}_{jk}$, and training input-output sequences x_1^T and p_1^T as following

$$\begin{aligned} B_{jk} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t} p_t x_t - \mu_{jk} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t} x_t}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t} x_t^2}, \\ \mu_{jk} &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t} p_t - \hat{B}_{jk} \sum_{n=1}^{N} \sum_{t=1}^{T_n} x_t}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}}, \\ \sigma_{jk}^2 &= \frac{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}(p_t - \hat{B}_{jk}x_t - \mu_{jk})^2}{\sum_{n=1}^{N} \sum_{t=1}^{T_n} \hat{\omega}_{jk,t}}. \end{aligned} \tag{3.29}$$

Next, new mixture probabilities can be estimated with posterior $\hat{\omega}_{jk,t}$ by solving the following maximization problem, which is a part of Equation 3.27 where the mixture probability is used.

$$\max_{\theta_2} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \sum_{k=1}^{M} \hat{\omega}_{jk,t} \log P(p_t | s_t = j, m_t = k, x_t; \theta_2). \tag{3.30}$$

Similar to the method for estimating a transition probability of our model, I use Lagrange multiplier and constraint $\sum_{k=1}^{M} P(m_t = k | s_t = j) = 1$ to solve for optimal $P(m_t = k | s_t = j)$ for all state j's and mixture m's.

$$P(m_t = k | s_t = j) = \frac{\sum_{t=1}^{T} \hat{\omega}_{jk,t}}{\sum_{l=1}^{K} \sum_{t=1}^{T} \hat{\omega}_{jl,t}} \tag{3.31}$$

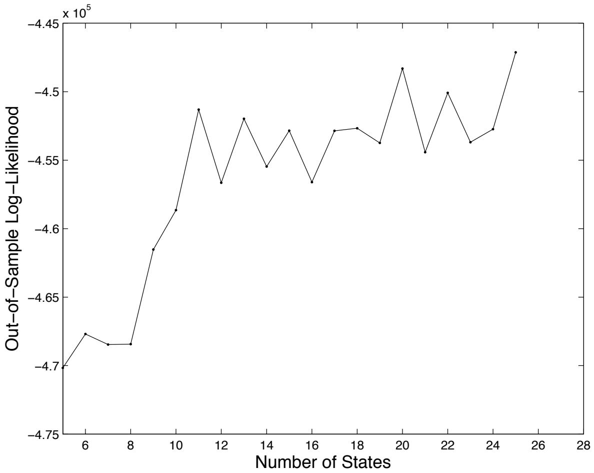


Figure 3-6: Change in out-of-sample log-likelihoods for stock AZN. Out of 675 trading days, 30% of them are randomly selected as out-of-sample and remaining 70% are used for training.

Initial State Probability

Lastly, estimating initial probability $P(s_1)$ for the model is the simplest of all. After estimating transition and output parameters $\hat{\theta}$ using methods described earlier, we compute initial probability as

$$P(s_1 = i) = P(s_1 = i | p_1^T, x_1^T; \hat{\theta}) = \hat{\gamma}_{1,i}. \tag{3.32}$$

Size of IOHMM

Using the supervised learning method discussed so far, I can train IOHMM price model. However, before discussing the performances of the model it is essential to address the issue of selecting size of IOHMM. Although I introduced a rigorous way of estimating model parameters of IOHMM with given number of mixtures and states, it is not so clear how many mixtures and states are appropriate to model prices. Unfortunately, there is no known method for solving model selection problem for IOHMMs with firm theoretical ground.

Therefore based on the empirical findings, I follow an Occam's Razor principle and identify the smallest possible size of IOHMM that I think is the good representation of the price process. For number of mixtures, I use 2 mixtures of Gaussians for each state to model asymmetric distribution of prices shown in Figure 3-3 and 3-4. For number of states, I tested how much improvement is made as a result of adding an extra state to the model in terms of improvement in out-of-sample likelihoods. Such test is done by first randomly selecting 30% of 675 trading days from our data as out-of-sample set. Let IOHMM(S) denote IOHMM with S number of states. I train IOHMM(S) with remaining 70% of the data, and compute the log-likelihood of 30% of out-of-sample data using the trained model. I do this for various number of states and plot the out-of-sample likelihoods over different number of states.

Figure 3-6 shows the change in out-of-sample log-likelihoods for stock AZN over number of states. From the figure, we no longer observe significant increase in out-of-sample log-likelihoods when $S > 12$. From this result I choose $S = 12$ for the stock AZN. The same method is applied to remaining 8 stocks and the resulting number of hidden states ranges from 10 to 16. The estimated parameters of price model for all 9 stocks are shown in Appendix A.2.

3.3.3 Price Generation

Once the IOHMM price model is trained with the actual sequence of price (p_t) and bid-ask spread (x_t) data, I use this model to generate a new sequence of price for the simulated orders. At the beginning of simulation, an initial state is randomly chosen using the initial state probabilities $P(s_1)$ as prior. Then for given state s_t and bid-ask spread x_t at time t, I generate a price from the following mixture of Gaussians

$$P(p_t|s_t = i, x_t) \quad = \quad \sum_{k=1}^{K} P(m_t = k|s_t = i) P(p_t|s_t = i, m_t = k, x_t), \tag{3.33}$$

where i is the state at time t and K is the number of mixture components where in our case $K = 2$. Once the price p_t is generated, I generate the next state s_{t+1} using the state transition probability $P(s_{t+1}|s_t)$.

Figure 3-7 shows PDFs of actual and simulated prices for the stock AZN. The sequence of bid-ask spread of length 2,328,429 from the actual data is used for simulation. To show how the price model reacts to the different sizes of bid-ask spread, I separately plot prices with bid-ask spread smaller (left) and greater (right) than 0.0025. The figure shows the price model successfully generates prices close to the actual data.

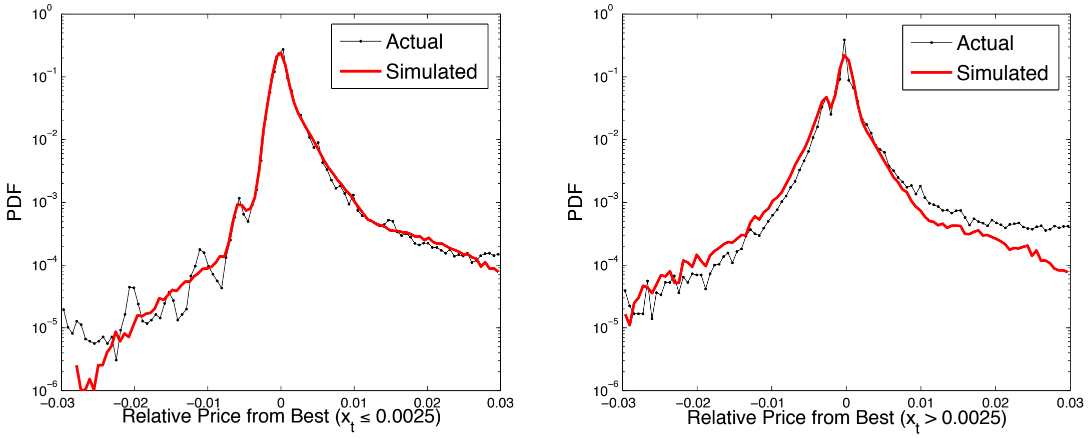


Figure 3-7: Empirical PDFs of the actual and the simulated price for the stock AZN. Total of 2,328,429 prices are simulated. The left (right) figure shows a price distribution when bid-ask spread x_t is $x_t \leq 0.0025$ ($x_t > 0.0025$).

3.4 Cancellation Time Model

The last component of the order flow model is a cancellation time model, which determines the duration of orders remaining in the order book. Every order has its cancellation time assigned by the cancellation time model. If an order gets executed before this duration, the cancellation time has no effect. The details of modeling method requires much space not feasible to put them in a single section. Thus in this section, I only describe an overview of the problem and the result. I explain the details of modeling method in Chapter 4.

The difficulty of modeling cancellation time is caused by missing data. Some orders are executed before they are canceled, therefore, their cancellation times are not known. In addition to this, since I do not have a full data describing the cancellation process, it is also not clear which statistical model to use and how to test it. In Chapter 4, I introduce a statistical technique, which helps overcoming the missing data problem and I also introduce a method indirectly verifying our estimated model.

The parametric form I introduce for the cancellation time model is called a q-Weibull distribution, which is a generalized form of a Weibull distribution. I will denote a PDF of cancellation time τ as $f_c(\tau)$ and its functional form has 3 parameters

$$f_c(\tau) \quad = \quad (2-q)\frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1} e_q^{-\left(\frac{\tau}{\lambda}\right)^{\alpha}}, \qquad (3.34)$$

where q-exponential function e_q is defined as $e_q^{-\tau} \equiv [1-(1-q)\tau]^{1/(1-q)}]$ if $1-(1-q)\tau \geq 0$ and $e_q^{-\tau} \equiv 0$ if $1-(1-q)\tau < 0$.

The estimated parameters of cancellation time model for all 9 stocks are shown in Appendix A.3. In Chapter 4, I introduce conditional variables, price and size of order, and study their effects on cancellation time. However, in order flow model, I use marginal distribution shown in Equation 3.34 to generate cancellation times since such effects are negligible. In Chapter 4, I also show that the order price has a very little effect on cancellation times and since the order flow model simulates orders with fixed size, the effect of order size on cancellation time can be safely ignored.

3.5 Discussion

This chapter introduces a stock order flow model consists of three sub-models – order sign, price, and cancellation time. Each sub-model simulates a component of order, and when they are combined, it simulates an order flow. The combining of these three components and results of simulations are discussed in Chapter 5.

In this chapter, I show how each component of order flow model can be modeled. The methods can be summarized as follows:

- The order signs are modeled with FARIMA model, which simulates buy/sell order signs forming a long memory transaction sign series. The estimated parameter of FARIMA model will ensure that actual and simulated transaction sign series to have similar Hurst exponents.

- The order price is modeled with IOHMM, which simulates a sequence of prices dependent on bid-ask spread. The result shows that IOHMM can successfully model asymmetry of price distribution and the dependence on bid-ask spread.

- The cancellation time is modeled with q-Weibull distribution. The details about the modeling method and results are not presented in this chapter. Chapter 4 is partly dedicated for this.

Chapter 4

Limit Order Removal Times Model

4.0 Preface

Part of this chapter is a continuation of modeling stock order flow described in Chapter 3.4. In this chapter I introduce removal time model comprised of limit order cancellation and execution time models. A cancellation time model I describe here is used in the order flow model.

4.1 Introduction

One of the widely used methods for trading financial securities is placing a limit order. A limit order is an order to transact a prespecified number of shares at a prespecified price, which gives the trader a control over the price at which the order is executed[1]. The primary advantage of a limit order is the absence of price risk by guaranteeing the execution price, but downside of it is the prespecified price can delay or even indefinitely prevent the execution. Therefore, when traders are faced with a decision placing a limit order, it is useful to know how much delay to expect.

In this chapter I propose a parametric model of limit order removal times, which can be used as a tool for understanding expected lifetime of limit orders in the market. The limit order execution times cannot be understood only with execution times from market data since limit order removals happen in two ways – execution and cancellation. The execution

[1] Here I denote limit order as effective limit order, which its prespecified price is not crossing the opposite best price (i.e. greater than best bid price when sell and lesser than best ask price when buy).

and the cancellation are two sides of same coin such that the observed execution times in the market data are results of executions occurred *before* the cancellation and vice versa. Thus to model execution times (time-to-fill) of limit orders, it is essential to also model cancellation times (time-to-cancel) of limit orders. Throughout this chapter I use the term "removal time" to refer to both time-to-fill and time-to-cancel.

The parametric form I introduce for the removal time model is a q-Weibull distribution, which is a generalized form of Weibull distribution. The q-Weibull distribution captures the diffusive behavior of the removal times studied by Eisler et al. (2007), which shows that limit order removal times decay asymptotically in time as a power law. Also I extend our approach to a conditional model and show that it captures similar effects of limit price and limit size on the limit order removal times studied by Lo et al. (2002). Several other studies explore the probability of limit order execution. Angel (1994), Hollifield et al. (1999), Battalio et al. (1999), Biais et al. (1995), Handa and Schwartz (1996), Harris and Hasbrouck (1996), Keim and Madhavan (1995), McInish and Wood (1986), and Petersen and Fialkowski (1994) are a few examples. Although none of these papers attempt to model removal times as I introduce in this paper, my model can be used as a tool to further study the issues they address.

4.1.1 Outline

The chapter is organized as follows: In Section 4.2 I identify problems of modeling limit order execution times when cancellation times are ignored and vice versa. In Section 4.3 I show that using a standard statistical technique to model removal times cannot accurately model the removal time's diffusive behavior and introduce a better approach of modeling it. In Section 4.4 I show the performance of the model by simulating removal times and comparing them with the empirical data. Then I extend the analysis to the conditional model and report additional findings. I conclude in Section 4.5.

4.2 Data Censoring and Survival Analysis

The survival analysis is a statistical analysis of death or failure time in biological or mechanical systems Cox and Oakes (1984); Kalbfleisch and Prentice (1980); Miller (1981). Since the goal is to model limit order removal times, which can be viewed as death or failure time of limit orders, I apply this technique. The advantage of a survival analysis is its ability to accommodate *censored observations*. For example, when modeling time-to-fill some orders are canceled before they are executed, therefore their execution times are “censored” due to their earlier cancellations. Although we do not know the time-to-fill of canceled orders, their time-to-cancel yields useful information – we know that they are not executed for at least a certain period of time. Despite the fact that the canceled orders do not have execution times, ignoring them can dramatically bias the time-to-fill estimation. A similar argument can be made when modeling the time-to-cancel of limit orders. In this case the executed orders are censored observations.

The survival analysis involves estimating the survival function $S(\tau)$, which is the inverse cumulative density function $F(\tau)$,

$$S(\tau) = 1 - F(\tau). \tag{4.1}$$

The survival function $S(\tau)$ can be estimated using either parametric or non-parametric methods. In both cases it will attempt to model the survival function using both censored and uncensored observations. The technical details of applying survival analysis on the data set will be explained in Section 4.3.

The amount of censored observations is directly related to how much we can benefit from the survival analysis. In this analysis such gains are inevitable since I am modeling time-to-fill and time-to-cancel simultaneously, which can be viewed as opposite sides of the same coin. When we have a data set with high cancellation rate, modeling time-to-cancel will not benefit much since the amount of censored observations is small, but modeling time-to-fill will benefit a lot due to high censoring rate and vice versa.

In the rest of this section I introduce a parametric form for time-to-fill and time-to-cancel models. Before applying the survival analysis I first focus on model parameter estimation only using the observed (uncensored) data and demonstrate the adaptiveness of the parametric form on the data set. Then I introduce removal time generation model which describes how the censoring occurs in our framework. Lastly I show ignoring the censored observations will lead to a biased removal time models, which makes applying survival analysis necessary.

4.2.1 Fitting Observed Data

One natural approach to modeling time-to-fill and time-to-cancel is to use the observed (uncensored) data. In this section I focus on choosing parametric form for the removal time models only using the observed data.

Exploring analysis of observed time-to-fill and time-to-cancel suggested the Weibull distribution as a possible candidate functional form, which is one of the most widely used distributions in survival data (Cox and Oakes, 1984). I found, however, that where the Weibull distribution fits the body well, it failed to fit the fat tails of the data. This led me to propose a more generalized parametric family called the q-Weibull.

The Weibull distribution is used to describe positive-valued random variables. The functional form of the PDF $f_w(x)$ of the Weibull distribution is given by Hallinan (1993); Johnson et al. (1994)

$$f_w(\tau) = \frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1} e^{-\left(\frac{\tau}{\lambda}\right)^{\alpha}}, \tag{4.2}$$

where $\alpha, \lambda > 0$. The survival and hazard rate of Weibull distribution $S_w(\tau)$ and $h_w(\tau)$ are defined as

$$S_w(\tau) = 1 - \int_0^{\infty} f_w(u)du = 1 - e^{-\left(\frac{\tau}{\lambda}\right)^{\alpha}}, \tag{4.3}$$

$$h_w(\tau) = \frac{f_w(\tau)}{1 - F_w(\tau)} = \frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1}. \tag{4.4}$$

The survival function $S(\tau)$ is the probability an event has not been removed after time τ and the hazard rate[2] is the estimate of relative risk which in this case is a probability of removal of an order relative to a probability of survival at time τ. Note that PDF, survival, hazard rate functions are different ways formulating the same removal time process and they all share the same parameters. From Equation 4.4, we can see that the Weibull distribution can have monotonically increasing, decreasing, or constant hazard rate when α is greater, lesser, or equal to 1. In the context of our analysis, an increasing (decreasing) hazard rate indicates orders being likely to be removed in a faster (slower) rate as τ increases.

[2] A hazard rate is also called a failure rate.

Because the observed time-to-fills and time-to-cancels are heavy tailed distributions I introduce a generalization of the Weibull distribution, called the q-Weibull distribution (Picoli et al., 2003). The extra parameter q gives an extra degree of freedom to fit a distribution with a tail that decays as a power law. The PDF of the q-Weibull distribution can be written as[3]

$$f_{qw}(\tau) \quad = \quad (2-q)\frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1} e_q^{-\left(\frac{\tau}{\lambda}\right)^{\alpha}}. \tag{4.5}$$

The PDF of q-Weibull distribution is derived from Weibull distribution in Equation 4.2 by replacing the exponential function with generalized q-exponential function.

The q-exponential function is defined as $e_q^{-\tau} \equiv [1-(1-q)\tau]^{1/(1-q)}]$ if $1-(1-q)\tau \geq 0$ and $e_q^{-\tau} \equiv 0$ if $1-(1-q)\tau < 0$. Also note that the generalization introduces two more parameter constraints: $q \neq 2$ and $\lambda(1/(1-q))^{1/\alpha} > \max(\tau)$ when $q < 1$. The q-exponential function is a generalization of the exponential function, that converges to it when $q = 1$ that arises from optimizing the Tsallis entropy (Tsallis, 1988; Gell-Mann and Tsallis, 2004). As the parameter q becomes greater than 1 the tail of the distribution gets fatter, and conversely when q becomes smaller than 1, the tail gets thinner. When the tail decays faster than the Weibull distribution ($q < 1$) the probability goes to zero for a finite value of τ due to its fast decay. Therefore, the newly introduced constraint forces all samples to have valid probabilities. Also, note that when we take the limit as $q \to 1$, the normal Boltzmann-Gibbs entropy is recovered and therefore, in our case, it will converge back to the normal Weibull distribution in Equation 4.2.

The survival function for the q-Weibull survival function is

$$S_{qw}(\tau) = 1 - \int_0^{\infty} f_{qw}(u)du = \left[1-(1-q)\left(\frac{\tau}{\lambda}\right)^{\alpha}\right]^{\frac{2-q}{1-q}}. \tag{4.6}$$

When the tail parameter $q > 1$ the q-Weibull distribution asymptotically converges to a power law of the form $P(\geq \tau) \propto \tau^{-\gamma}$ for large τ. The tail exponent γ can be written in terms of q and α as $\gamma = \alpha \cdot (q-2)/(1-q)$ when $q > 1$.

The q-Weibull distribution is a generalized form of several other distributions. As already mentioned, in the limit $q \to 1$ the q-Weibull distribution converges to the Weibull distribution. Taking the limit as $\alpha \to 1$ reduces to a q-exponential. Taking the limits $q \to 1$

[3] The $(2-q)$ in $f_{qw}(\tau)$ is to satisfy $\int_0^{\infty} f_{qw}(u)du = 1$.

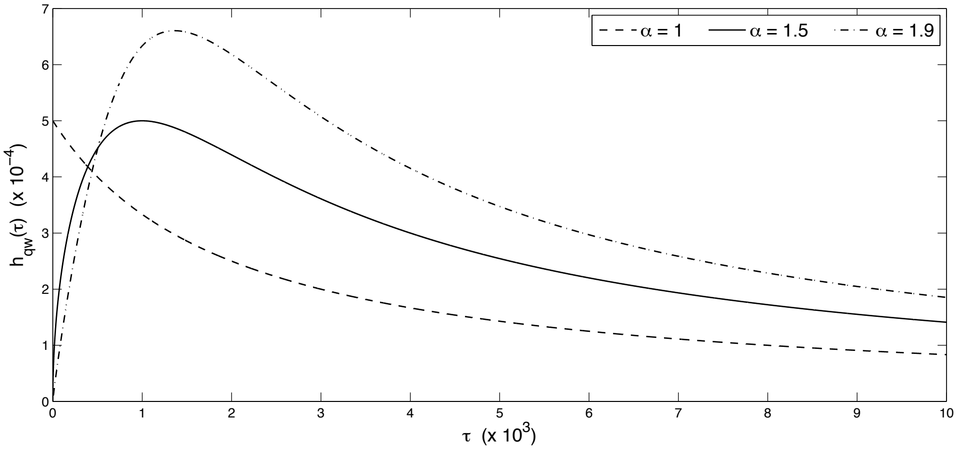


Figure 4-1: Different shapes of the hazard function $h_{qw}(\tau)$ for the q-Weibull distribution. The values for q and λ are fixed to 1.5 and 1000 respectively. The α values are varied – 1, 1.5, and 1.9

and $\alpha \to 1$ simultaneously gives an exponential distribution, and taking the limit as $q \to 1$, $\alpha \to 2$, and $\lambda \to \sqrt{2}\beta$ together gives a Rayleigh distribution.

The hazard function for the q-Weibull distribution, $h_{qw}(\tau)$, is defined as

$$h_{qw}(\tau) = \frac{f_{qw}(\tau)}{1 - F_{qw}(\tau)} = \frac{(2-q)\frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1}}{1-(1-q)\left(\frac{\tau}{\lambda}\right)^{\alpha}} \ . \tag{4.7}$$

Unlike Weibull distribution the hazard function for the q-Weibull distribution can be non-monotonic when $\alpha > 1$ and $q > 1$. Figure 6-6 shows the different shapes of q-Weibull hazard functions when α is varied. When $\alpha > 1$ and $q > 1$, the hazard rate function is non-monotonic.

Figure 4-2 shows Weibull and q-Weibull fits to the observed time-to-fills and time-to-cancels of stock AZN. The survival plots are shown in log-log scale to emphasize the tail behavior of removal times[4]. As shown in the figure the empirical survival plot shows slower decay than normal Weibull can model resulting a poor fit towards large τ. In contrast q-Weibull introduces an extra parameter q, which controls the fatness of the tail resulting a better fit.

Next we need to check if the resulting models are good models for the underlying removal time processes. By ignoring the censored data, one critical fact I ignored is that time-to-fill process is a censoring process of time-to-cancel process and vice versa. Do these models

[4] I refer all removal times in event times.

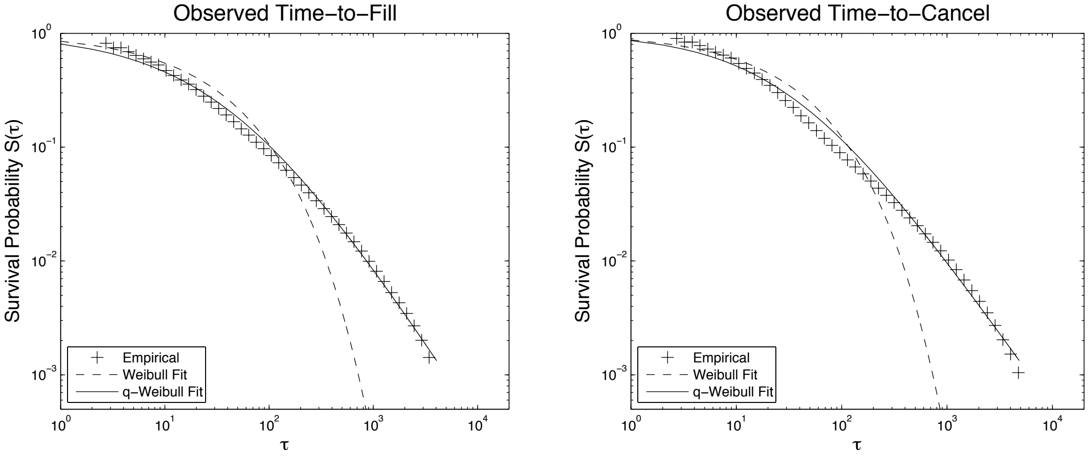


Figure 4-2: Weibull and q-Weibull fits to the empirical survival probabilities of observed time-to-fills (left) and observed time-to-cancels (right) of stock AZN. The censored observations are ignored. For the observed time-to-fill estimated Weibull parameters are $\alpha = 0.57$ and $\lambda = 24.18$, and estimated q-Weibull parameters are $q = 1.30$, $\alpha = 0.62$ and $\lambda = 6.30$. For the observed time-to-cancel estimated Weibull parameters are $\alpha = 0.60$ and $\lambda = 29.01$, and estimated q-Weibull parameters are $q = 1.35$, $\alpha = 0.71$ and $\lambda = 7.53$.

really describe the distributions of time-to-fill and time-to-cancel of limit orders? To test this, I introduce removal time generation model describing how the censoring occurs in this framework. Then using the removal time generation model I can test if my estimated time-to-fill and time-to-cancel models simulate observed removal times similar to the empirical data.

4.2.2 τ-Generation Model

In this subsection I introduce a removal time generation model describing how censoring occurs. I define the time-to-fill of an order to be the time it takes for an order to be fully executed in the market. Market participants who place a limit order are allowed to cancel it at anytime as long as the order is not executed. In the data set an average of 72% of limit orders are canceled before execution. In order to understand how time-to-fill and time-to-cancel are observed in the data set, it is important to model an underlying mechanism of time-to-fill and time-to-cancel, which we call the *τ-generation model.*

Let's define τ_i^{obs} as the observed time[5] order i took to get either executed or canceled in the order book. All orders will eventually be removed from the book either by execution

[5] The removal times are measured in event time.

or cancellation. Although we cannot always directly observe these, I define two underlying distributions $P_e(\tau)$ and $P_c(\tau)$, which determine the order i's execution and cancellation times, τ_i^e and τ_i^c.

When the order is placed, the market will determine when the order will be executed according to $P_e(\tau)$ and the market participants will decide when the order will be canceled according to $P_c(\tau)$. From this mechanism, the orders will either be executed or canceled whichever one occurs first. From the data set we can also see if order i was canceled or executed, therefore, we introduced an indicator function $I_e(i)$ which outputs 1 for the executed order and 0 for the canceled order. In summary, the model that generates *observed* time-to-fill or *observed* time-to-cancel of an order i is as follows[6]:

$$\begin{aligned} \tau_i^c \sim P_c(\tau), &\quad \tau_i^e \sim P_e(\tau) \\ \tau_i^{obs} = min(\tau_i^c, \tau_i^e), &\quad I_e(i) = \{0, 1\} \end{aligned} \tag{4.8}$$

where the underlying removal time distributions $P_e(\tau)$ and $P_c(\tau)$ are unknown, and τ_i^{obs} and $I_e(i)$ are known from the data.

Now we can simulate the τ-generation process of Equation 4.8 with previously estimated removal time models using q-Weibull distributions, $\hat{P}_e(\tau)$ and $\hat{P}_c(\tau)$. In section 4.2.1, I ignored the censored data and assumed that $\hat{P}_e(\tau) = P(\tau_i^{obs}|I_e(i) = 1)$ and $\hat{P}_c(\tau) = P(\tau_i^{obs}|I_e(i) = 0)$. If censored data does not bias the underlying removal time processes too much, the simulated observed time-to-fills and time-to-cancels by τ-generation model will form similar distributions to the empirical ones.

To simulate removal times for N orders, we sampled time-to-fill and time-to-cancel values, τ_i^e and τ_i^c, for $i = 1...N$ from estimated removal time distributions, $\hat{P}_e(\tau)$ and $\hat{P}_c(\tau)$. Once a $(\tau_i^e,\ \tau_i^c)$ pair is generated for an order i, we take the minimum of the two as the observed removal time τ_i^{obs} of an order i, and set the value of indicator function $I_e(i)$ to be 1 when τ_i^e is picked and 0 otherwise.

Figure 4-3 shows the survival plots of empirical and simulated observed time-to-fills and time-to-cancels. For simulated removal times 1,843,713 orders are simulated, which is equal to the number of orders in the empirical data. The figure shows after censoring mechanism is introduced by the τ-generation model, the simulated observed removal time for both

[6] The symbol "$\sim$" in Equation 4.8 means "sampled from" a given conditional distribution.

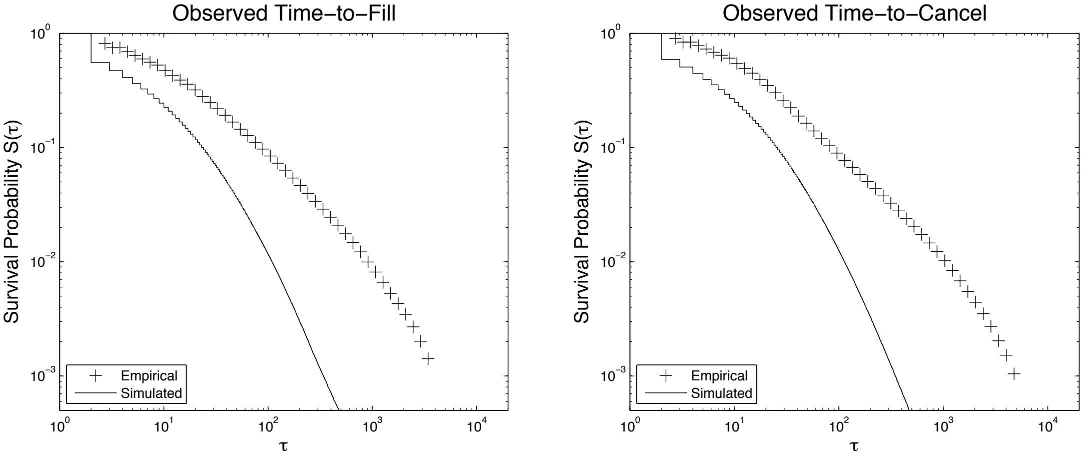


Figure 4-3: Survival plots of actual vs. simulated observed removal times for stock AZN: observed time-to-fills (left) and observed time-to-cancels (right). For simulated observed removal times q-Weibull distribution is used only using the uncensored data.

time-to-fill and time-to-cancel have drastically different distributions to the empirical ones. This indicates ignoring censored data during estimation will lead to biased removal time processes and therefore, applying survival analysis technique is necessary.

In the next section I introduce technical details of various model estimation methods using the survival analysis technique with q-Weibull distributions.

4.3 Estimation Methods

As I described in discussing the τ-generation model in section 4.2, the market data does not contain full information about removal times – if an order is executed, its time-to-cancel is unknown and if an order is canceled its time-to-fill is unknown. I have shown that these censored data can lead to biased models. For instance when modeling underlying time-to-fill process, some limit orders are canceled before execution and more persistent orders are much more likely to be canceled, which biases the time-to-fill distribution towards shorter execution times.

In this section, I introduce a survival analysis technique that deals with cases like this. I first discuss two ways of applying survival analysis in modeling limit order removal times using q-Weibull distributions – maximum likelihood estimation, and least squares fitting of Kaplan-Meier estimators. However, I find that when we are using to model fat-tailed distributions such as q-Weibull, this technique does not work well when the number of censored observations dominates the number of uncensored observations. I introduce a third method of resolving this issue.

The examples of modeling methods will be shown for time-to-fill distribution for the stock AZN. The same methods can be applied to other stock data or to model time-to-cancel process.

4.3.1 Maximum Likelihood Estimation (MLE)

Under the parametric method the survival analysis defines likelihood function $L_e(\theta)$ for time-to-fill with the q-Weibull distribution as

$$L_e(\theta) = \prod_{i=1}^{N} f_{qw}(\tau_i;\theta)^{I_e(i)} S_{qw}(\tau_i;\theta)^{1-I_e(i)}, \tag{4.9}$$

where θ is the set of q-Weibull parameters $\{q, \alpha, \lambda\}$, N is the number of orders, τ_i is the removal time of order i, and $I_e(i)$ is the indicator function having 1 for executed orders and 0 for canceled orders.

Unlike ordinary MLE technique our likelihood function $L_e(\theta)$ in Equation 4.9 not only consists of products of probability densities $f_{qw}(\tau)$ of the time-to-fill of executed orders, but also products of survival probabilities $S_{qw}(\tau)$ of the time-to-cancel of canceled orders, which are uncensored and censored observations respectively. Also note that $f_{qw}(\tau)$ and $S_{qw}(\tau)$ share the same q-Weibull parameters θ. By maximizing the above likelihood $L_e(\theta)$ we can obtain a time-to-fill model, which can correctly simulate potential time-to-fill of every orders so that some orders will remain (survive) in the order book long enough for the time-to-cancel model to identify them as canceled orders.

The likelihood for the time-to-cancel model $L_c(\theta)$ is the same as Equation 4.9, except the time-to-cancel and the time-to-fill are applied to maximize probability densities and survival probabilities respectively by swapping $I_e(i)$ and $1 - I_e(i)$.

When estimating parameters using MLE it is more convenient to maximize the log-likelihood function for q-Weibull,

$$\begin{aligned} l_e(q,\alpha,\lambda) \quad = \quad & \sum_{i=1}^{N} \log\left((2-q)\frac{\alpha}{\lambda}\left(\frac{\tau_i}{\lambda}\right)^{\alpha-1} e_q^{-\left(\frac{\tau_i}{\lambda}\right)^{\alpha}}\right)^{I_e(i)} + \\ & \sum_{i=1}^{N} \log\left(\left[1-(1-q)\left(\frac{\tau_i}{\lambda}\right)^{\alpha}\right]^{\frac{2-q}{1-q}}\right)^{1-I_e(i)} . \end{aligned} \tag{4.10}$$

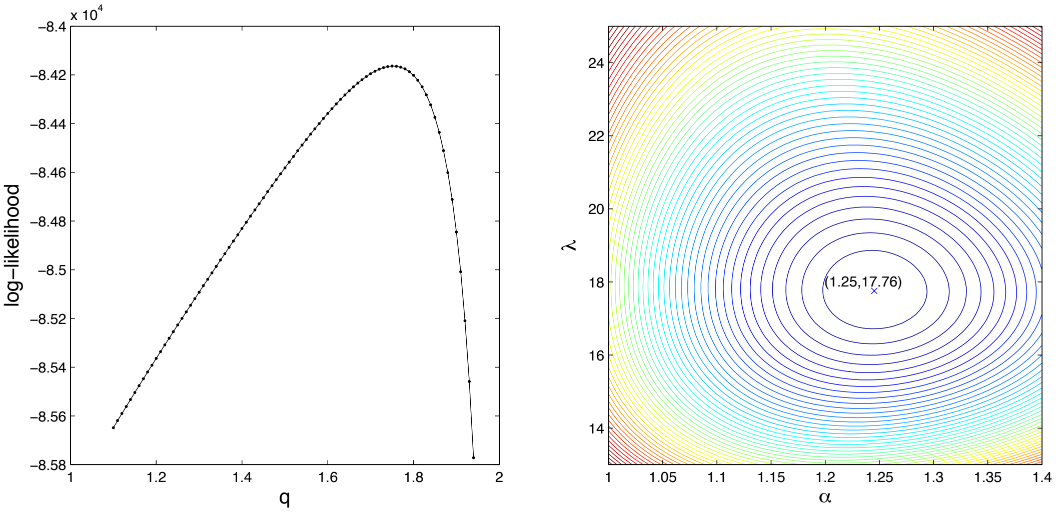


Figure 4-4: q-Weibull MLE estimation for $P_e(\tau)$ for the stock AZN. Change in log-likelihood $l'_e(\alpha, \lambda|q)$ given different values of q. The optimal value for q, which results the maximum log-likelihood is $q^* = 1.75$ (left). Contour plot of log-likelihood $l'_e(\alpha, \lambda|q^* = 1.75)$ of maximization problem (eqn 4.10) given optimal q^* (right). The optimal α and λ are $(\alpha^*, \lambda^*) = (1.25, 17.76)$. In this data set, 77.49% of orders are canceled.

Finding an optimal set of parameters that maximizes the Equation 4.10 is difficult since the log-likelihood function is not convex, which results in multiple local maxima. To simplify the computation we fix q and perform the optimization over α and λ only, which results in a unique maximum, then exhaustively search over q to find the values $(q^*, \alpha^*, \lambda^*)$ with the maximum log-likelihood.

Figure 4-4 shows the result of MLE algorithm described above for the time-to-fill distribution $P_e(\tau)$ for the stock AZN. Since we estimate time-to-fill, cancellation times are used as censored observations. The left figure plots total log-likelihoods of the AZN data set for various q values and the right figure is a contour plot of log-likelihoods given an optimal q^*, illustrating that there is a unique maximum for fixed q.

4.3.2 Kaplan-Meier Estimators and Least Squares Fit

Another way of applying survival analysis is to use a non-parametric estimator using the statistical technique developed by Kaplan and Meier (1958). By using the product limit method, Kaplan and Meier developed a non-parametric maximum likelihood estimate of

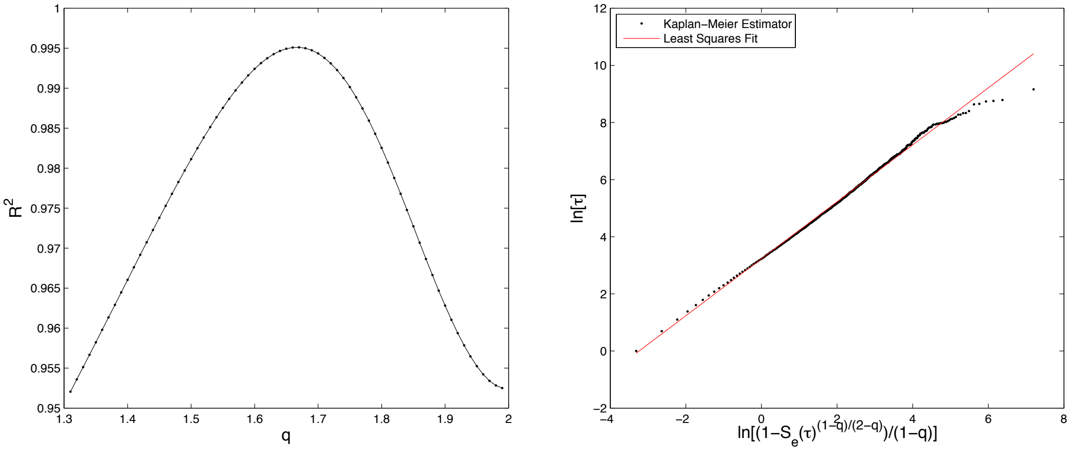


Figure 4-5: q-Weibull least squares fit for $P_e(\tau)$ of stock AZN: The change in R^2 over different q's (left). The optimal value q^* is 1.67. The plot of Equation 4.12 for q^* and its least squares fit (right). The optimal α and λ are $(\alpha^*, \lambda^*) = (1.00, 25.23)$. In this data set, 77.49% of orders are canceled.

$\hat{S}_e(\tau)$ of the form

$$\hat{S}_e(\tau) = \prod_{\tau_j \leq \tau} \left(1 - \frac{n_j - d_j}{n_j}\right). \tag{4.11}$$

To estimate time-to-fill, τ_j is the time until j executions are observed, n_j is the number of orders that are remaining without execution or cancellation prior to time τ_j, and d_j is the number of executions at time τ_j. This treats canceled orders as censored observations. Similarly a non-parametric estimate for the time-to-cancel survival function can be obtained by taking executed orders as censored observations.

Once we obtain a non-parametric estimator of the survival function I fit it with the q-Weibull survival function using least squares. This is conveniently done by rearranging the q-Weibull survival function in Equation 4.6 into linear form and setting $S_{qw}(\tau) = \hat{S}_e(\tau)$ and $\tau = \hat{\tau}$, where $\hat{S}_e(\tau)$ is the Kaplan-Meier estimate for $\hat{\tau}$.

$$\ln(\hat{\tau}) \quad = \quad \frac{1}{\alpha} \cdot \ln\left(\frac{1 - \hat{S}_e(\tau)^{\frac{1-q}{2-q}}}{1-q}\right) + \ln(\lambda). \tag{4.12}$$

In order to apply the least squares method to Equation 4.12, q needs to be known. Therefore, as I did for the MLE estimation, I fix q and perform an exhaustive search. For the optimal q^* the plot for Equation 4.12 will be the straightest line, with slope $1/\alpha^*$ and intercept $\ln(\lambda^*)$. The optimal q^* is found by maximizing R^2 over q.

Figure 4-5 shows the result using this method. The left figure shows the change of R^2 over different q's, and the right figure compares Equation 4.12 for q^* to the Kaplan-Meier estimator.

4.3.3 Inconsistency with Estimation Methods and Improvement

Although two estimation methods introduced are applied to the same data, the q-Weibull parameter estimates significantly differs from one another. The estimated q-Weibull parameters $\{q^*, \alpha^*, \lambda^*\}$ for time-to-fill model for AZN are $\{1.75, 1.25, 17.76\}$ for MLE, and $\{1.67, 1.00, 25.23\}$ for least squares fit to Kaplan-Meier estimators. Moreover, due to censoring, we cannot directly compare the results to the true underlying time-to-fill distribution, which leads to having no direct way of finding which of these two are the better estimates. A possible reason that leads to such inconsistency is high censoring rate. The set of AZN orders we tested has 77.49% of canceled orders and since I am estimating time-to-fill model in the previous examples, canceled orders are censored observations.

The censoring in our data set is not a uniform random event. For both time-to-fill and time-to-cancel models, when τ increases the probability of an observation becoming a censored one will increase. This will cause relatively larger loss of information on the tail of the distribution than the body. Since the q-Weibull has a parameter q which is sensitive to the tail, we cannot accurately estimate q when the censor rate is too high.

I propose an alternative estimation method that works better for high censoring rates. When the data set has a high cancellation rate we can use canceled orders to model cancellation and then using the resulting model as a parameter model for the censoring process for time-to-fill.

Define the PDF function for observed time-to-fill, $P(\tau_i^{obs}|I_e(i) = 1)$, as $f_e^{obs}(\tau)$. Then we can formulate $f_e^{obs}(\tau)$ in terms of time-to-fill and time-to-cancel distributions,

$$f_e^{obs}(\tau) = \frac{f_e(\tau)(1 - F_c(\tau))}{\int_0^\infty f_e(u)(1 - F_c(u))du} = \frac{f_e(\tau)S_c(\tau)}{\int_0^\infty f_e(u)S_c(u)du} . \tag{4.13}$$

The τ-generation model explains why the equality in Equation 4.13 holds. For order i, τ_i^e and τ_i^c are sampled from distributions $f_e(\tau)$ and $f_c(\tau)$. If $\tau_i^e < \tau_i^c$ the order will be executed after τ_i^e and if $\tau_i^e > \tau_i^c$ it will be canceled after τ_i^c. Thus the probability that order i is executed after time τ_i^e is $f_e(\tau_i^e) \cdot \int_{\tau_i^e}^{\infty} f_c(u)du$ – the probability of τ_i^e is sampled from $f_e(\tau)$ multiplied by the probability that a value greater than τ_i^e will be sampled from $f_c(\tau)$. Then $f_e^{obs}(\tau)$ will be the normalized distribution of the above probabilities for all τ's. For observed time-to-cancel distribution $f_c^{obs}(\tau)$, the same equality condition will hold by swapping time-to-fill and time-to-cancel distributions in Equation 4.13. Also note that $\int_0^{\infty} f_e(u)S_c(u)du$ is the average probability of execution r_e, which can be directly measured from the data set.

We can use Equation 4.13 to yield a better time-to-fill model by the following: If the majority of orders in data set are canceled (censored) orders, we can find a fairly accurate q-Weibull estimate for the time-to-cancel model using the Kaplan-Meier estimator with least squares fitting as we explained in section 4.3.2. The observed time-to-fill in the data set $f_e^{obs}(\tau)$ can be estimated with q-Weibull distribution. Given the empirical execution rate $\hat{r}_e$, the Kaplan-Meier estimators for the time-to-cancel survival function $\hat{S}_c(\tau)$, and the PDF of observed time-to-fill from the data set $\hat{f}_e^{obs}(\tau)$, the PDF of underlying time-to-fill $f_e(\tau)$ can be solved with the following Equation

$$f_e(\tau) \quad = \quad \frac{\hat{f}_e^{obs}(\tau) \cdot \hat{r}_e}{\hat{S}_c(\tau)}. \tag{4.14}$$

Once we run the above method, Equation 4.14 will yield set of $(\tau, f_e(\tau))$ pairs, which are non-parametric estimators of $f_e(\tau)$. Similar to what we did in section 4.3.2 with Kaplan-Meier estimators, we can apply q-Weibull least squares fit on these set of non-parametric estimators to find q-Weibull parameters for time-to-fill model. When there are more executed orders than canceled orders, we can apply a similar method to model time-to-cancel by swapping time-to-fill and time-to-cancel in Equation 4.14.

To compare the estimation methods I test them on synthetic data generated by the τ-generation model. Several sets of arbitrary q-Weibull parameters for $P_e(\tau)$ and $P_c(\tau)$ are picked to vary the cancellation rate. Since the time-to-fill and time-to-cancel models are known we can compare the performance of the estimation methods.

To simulate N orders I based on an assumption that observed removal times are emerged from the τ-generation model I described in Section 4.2.2. I sampled time-to-fill and time-to-cancel values, τ_i^e and τ_i^c, for $i = 1...N$ from known distributions, $P_e(\tau)$ and $P_c(\tau)$. Once a (τ_i^e, τ_i^c) pair is generated for an order i, we take the minimum of the two as the observed

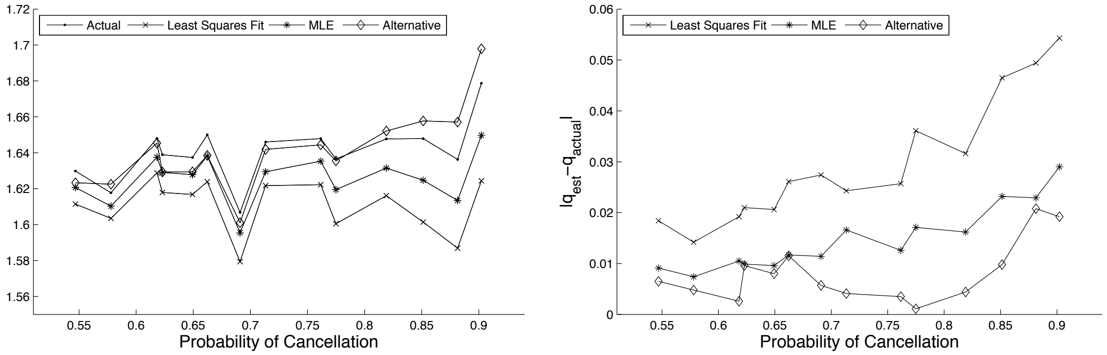


Figure 4-6: Comparison of q-Weibull tail parameter estimates of synthetic data sets with various cancellation probabilities. The survival analysis was used to exploit the censored observations. For each cancellation probability, 100 data sets with 10,000 orders are used to calculate the mean of tail estimates. The left plot shows q vs probability of cancellation for actual, least squares fit of Kaplan-Meier estimators, MLE, and Alternative estimation methods. The right plot shows the absolute difference between actual and estimated tail parameters of the three estimation methods for various probabilities of cancellations.

removal time τ_i^{obs} of an order i, and set the value of indicator function $I_e(i)$ to be 1 when τ_i^e is picked and 0 otherwise. For each data set 10,000 orders are generated.

Using the synthetic data I estimated q-Weibull parameters for the time-to-fill model using three estimation methods. I repeated data generation and estimation 100 times for each parameter set. In Figure 4-6 the plot on the left compares the actual and mean of estimated values of q for all three estimation methods and for various probabilities of cancellation. A larger probability of cancellation implies a larger number of censored observations in the synthetic data set. The result in the figure shows that as the number of censored observation increases it is more difficult to estimate the tail parameter q. Also it shows the alternative estimation method introduced produces a better tail estimates than the other two methods. The left plot shows the difference in tail estimates more clearly. The y-axis of the plot is the absolute difference between the actual and mean of estimated tail parameters. It shows that the tail estimates from the alternative method are consistently better than the other two estimates.

The amount of inaccuracy of tail parameter estimates shown here affects the overall distributions significantly. Figure 4-7 is a plot of the absolute difference between the actual and estimated survival functions $|S_{est}(\tau) - S_{actual}(\tau)|$ over various τ's for synthetic data with cancellation probability of 0.90. The result shows that the alternative method estimated

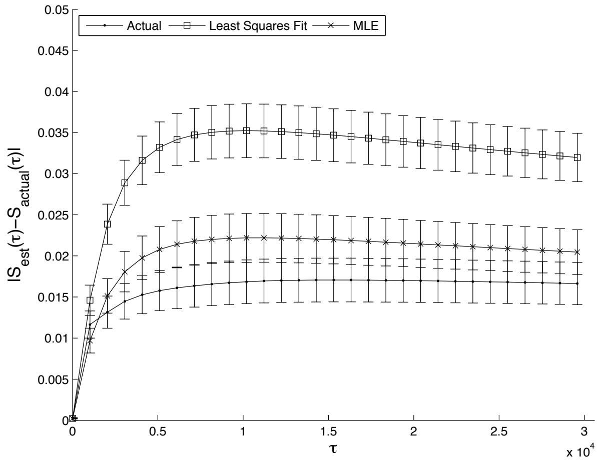


Figure 4-7: Absolute difference between actual and estimated survival functions $|S_{est}(\tau) - S_{actual}(\tau)|$ using three different estimation methods. Arbitrary set of q-Weibull parameters for time-to-fill and time-to-cancel are used to generate orders with cancellation probability of 0.9. The 95% confidence intervals are calculated using 100 generations of synthetic data set with 10,000 orders in each set.

significantly more accurate estimates when the fraction of censored observations is high.

4.4 Empirical Analysis

I now turn to the empirical analysis of limit order removal time data of 9 stocks in LSE using our τ-generation model. I first focus on unconditional removal time model, and then extend this to conditional model by introducing conditional variables. I show that introduced survival analysis technique can successfully model underlying removal time processes, which simulate observed time-to-fills and time-to-cancels close to the empirical distributions. Then I discuss the effects of introduced conditional variables on the limit order removal times, and economic implications of change in tail exponent of removal time distributions over conditional variables. Lastly I compare our modeling method to the previously known method for modeling time-to-fill using first passage time analysis.

4.4.1 Unconditional Model: With vs. Without Censored Data

As shown in Figure 4-3 ignoring censored observations in modeling time-to-fill and time-to-cancel processes as $\hat{P}_e(\tau) = P(\tau_i^{obs}|I_e(i) = 1)$ and $\hat{P}_c(\tau) = P(\tau_i^{obs}|I_e(i) = 0)$ results inaccurate models, which fail to simulate observed removal time distributions close to the

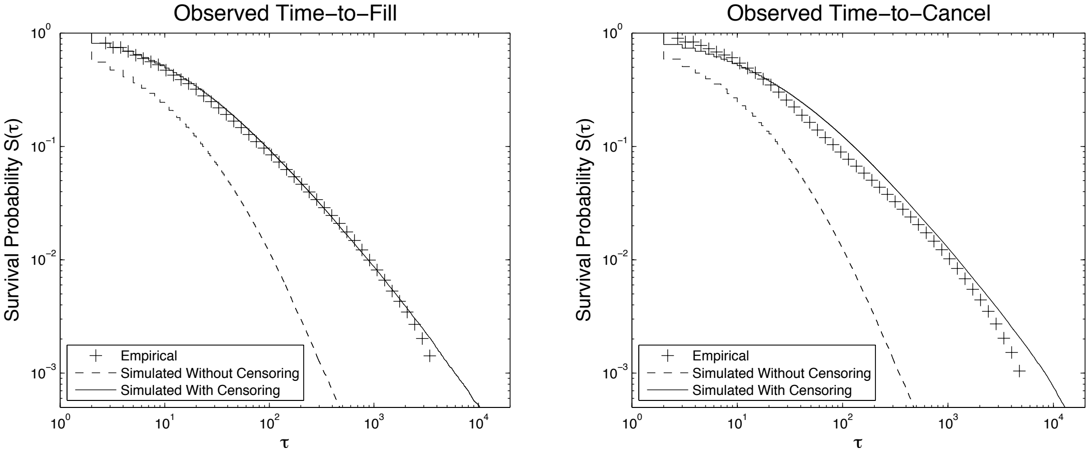


Figure 4-8: Survival plots of empirical vs. simulated observed removal times of stock AZN with and without censored data: observed time-to-fills (left) and observed time-to-cancels (right). The q-Weibull distribution is used for both removal time models. When modeling removal times with censored data, 69.18% of them are canceled orders, therefore, time-to-fill model is indirectly modeled deriving from the estimate of time-to-cancel model and the observed time-to-fill distribution as we described in Section 4.3.3.

empirical ones. An alternative modeling method using the survival analysis is introduced in Section 4.3 with detailed estimation method to overcome the difficulty of modeling fat tailed distribution when the number of censored observations dominates the number of uncensored observations.

Figure 4-8 illustrates how much the removal time models are improved when censored data are included in the estimation. In stock AZN 69.18% of orders are canceled orders, which the number of censored dominates the number of uncensored observations when modeling time-to-fill. Therefore in estimating time-to-cancel we use the least square fit of Kaplan-Meier estimators described in Section 4.3.2, and in estimating time-to-fill we derive the parameter estimates from the time-to-cancel model and the observed time-to-fill distribution as described in Section 4.3.3. In simulating both observed time-to-fill and time-to-cancel the new method results significantly better removal time models than the ones modeled without the censored data. The same analyses are done with rest of the 8 stocks[7]. Table 4.1 and 4.2 report estimated q-Weibull parameters for time-to-fill and time-to-cancel respectively. Also using the parameter estimates, Figure 4-9 and 4-10 show the estimated removal time models simulate observed time-to-fills and time-to-cancels close to the empirical distributions for all stocks.

[7] Other 8 stocks we analyzed are BLT, BSY, LLOY, PRU, RTO, RTR, TSCO, and VOD.

Stock	q	α	λ
AZN	1.80	0.9459	6.0002
BLT	1.73	0.8483	7.4387
BSY	1.79	0.9430	4.4881
LLOY	1.81	1.0275	3.6813
PRU	1.80	0.9804	4.5808
RTO	1.76	0.9423	5.2043
RTR	1.80	0.9970	3.6883
TSCO	1.77	0.9655	4.2397
VOD	1.06	0.3241	86.0865

Table 4.1: q-Weibull parameter estimates of execution time (time-to-fill) model for the 9 stocks.

Stock	q	α	λ
AZN	1.43	0.7024	9.3382
BLT	1.48	0.8020	6.4889
BSY	1.43	0.7197	9.8304
LLOY	1.38	0.6461	15.1451
PRU	1.50	0.8786	8.6154
RTO	1.38	0.6637	9.5057
RTR	1.40	0.6534	8.5210
TSCO	1.32	0.5983	22.1437
VOD	1.09	0.4153	118.5829

Table 4.2: q-Weibull parameter estimates of cancellation time (time-to-cancel) model for the 9 stocks.

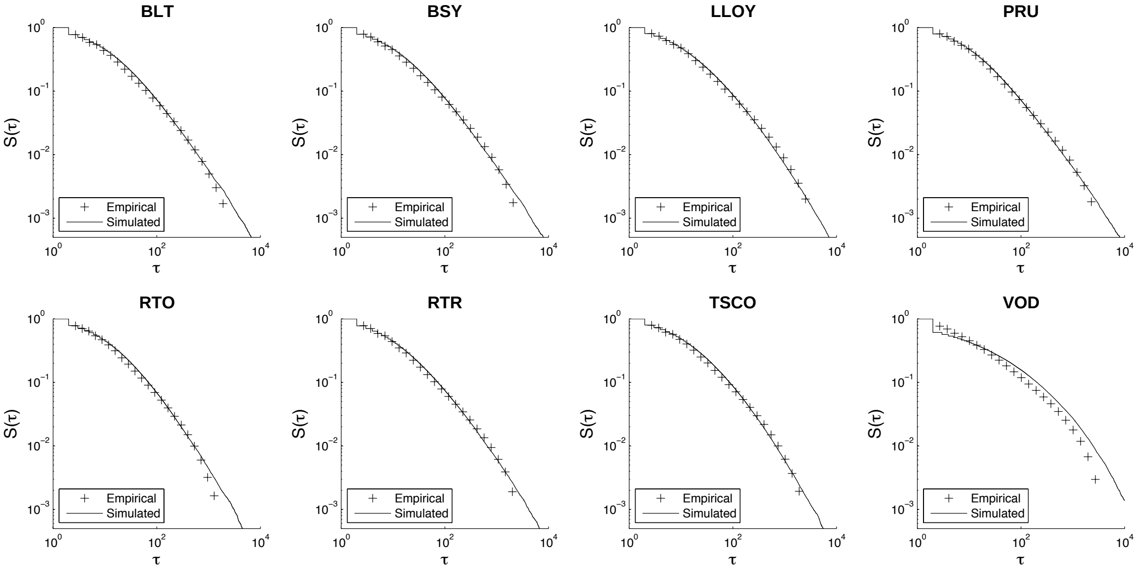


Figure 4-9: Survival plots of empirical vs. simulated *observed time-to-fills* of 8 stocks.

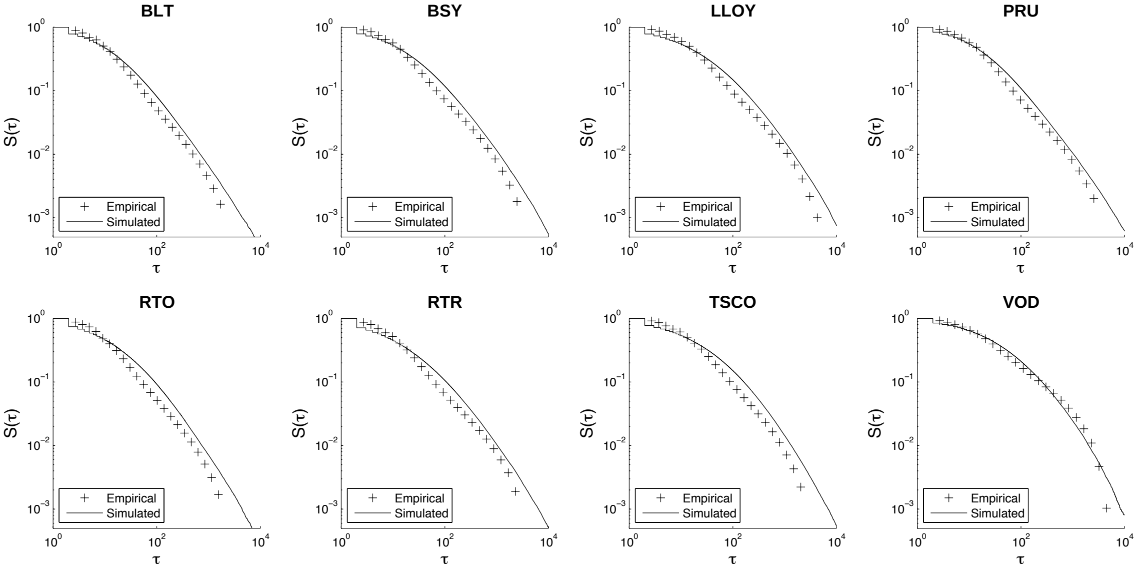


Figure 4-10: Survival plots of empirical vs. simulated *observed time-to-cancels* of 8 stocks.

4.4.2 Conditional Model

I now extend the estimation method to model conditional removal time model. There are many micro and macro economic variables that can be introduced to model their effects on limit order removal times. In this section I study two variables which are directly related to the limit order conditions, the price and size of a limit order.

Conditional Variables

In order to understand how the removal times are affected by order size and limit price I introduce two conditional variables s and Δp and model underlying conditional distributions[8] $P_e(\tau|\cdot)$ and $P_c(\tau|\cdot)$ based on an assumption that observed removal times are generated from τ-generation model.

The size of an order s is represented in monetary terms since the number of shares has a different meaning depending on price. The size is defined as

$$s = p_f \times (\# \text{ of shares}), \tag{4.15}$$

where p_f is the limit price.

The mid-point price is defined as the average of the logarithm of the best bid and ask prices $p_{mid} = (\log(p_{bid}) + \log(p_{ask}))/2$ at the time of limit order submission. To see the effect of the limit order price on removal times we use the distance between the order's limit price and the mid-point price Δp, defined as

$$\Delta p = \begin{cases} p_{mid} - \log(p_f) & \text{if buy order} \\ \log(p_f) - p_{mid} & \text{if sell order.} \end{cases} \tag{4.16}$$

Larger Δp corresponds to less aggressive order placement and larger expected order removal time. Similarly, as the size of order s increases we expect larger expected order removal times.

The parameters for the conditional distributions of order removal times are estimated by dividing the data set into bins such that each bin will contain orders with certain ranges of conditional values. The bins containing the largest s's and Δp's are discarded in the analysis since the variance of conditional values are too large. Using each subset we estimate

[8] The "$\cdot$" in $P_e(\tau\ |\cdot)$ and $P_c(\tau\ |\cdot)$ represents either s or Δp

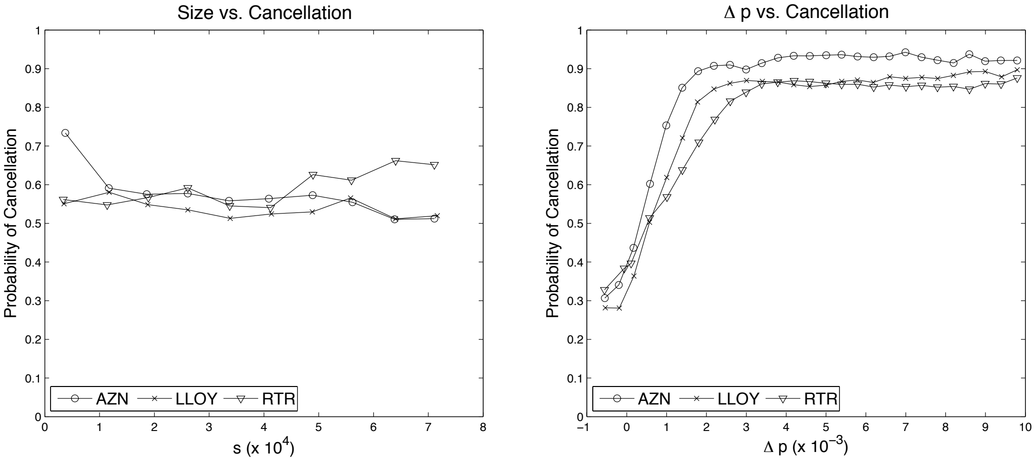


Figure 4-11: Change in probabilities of cancellation of limit orders given s and Δp: Mean of s vs. Cancellation (left), Mean of Δp vs. Cancellation (right). Three are picked from the ten stocks used in our analysis to plot the figure. The other seven stocks follow similar patterns to those are shown here.

parameters for conditional distributions for time-to-cancel, $P_c(\tau|s)$ and $P_c(\tau|\Delta p)$, and time-to-fill, $P_e(\tau|s)$ and $P_e(\tau|\Delta p)$.

When the data set is divided into bins based on conditionals, the proportions of executed and canceled orders can be dramatically shifted. For instance, the 69.18% of entire AZN data are canceled orders, but when the data set is divided based on Δp the amount of canceled orders ranges from 30% to 90%. Figure 4-11 shows cancellation probabilities of orders as a function of mean of Δp and s in each bin for 3 stocks among the 9 stocks we studied. The remaining 6 stocks show similar effects. The figure shows that the probability of cancellation does not change on s (left) but increases as Δp gets larger (right).

From this result one may conclude the patterns of market participants' order cancellation behavior such that Δp plays an dominant role when traders make cancellation decisions. This is an incorrect observation since we do not know how time-to-fill model will change as Δp increases. If expected time-to-fill dramatically increases over Δp, we may see the same result even with the absence of change in traders' cancellation behavior. Therefore, in order to understand the conditional effects on both time-to-fill and time-to-cancel, it is important to correctly model the τ-generation process from market data using estimation methods described in Section 4.3.

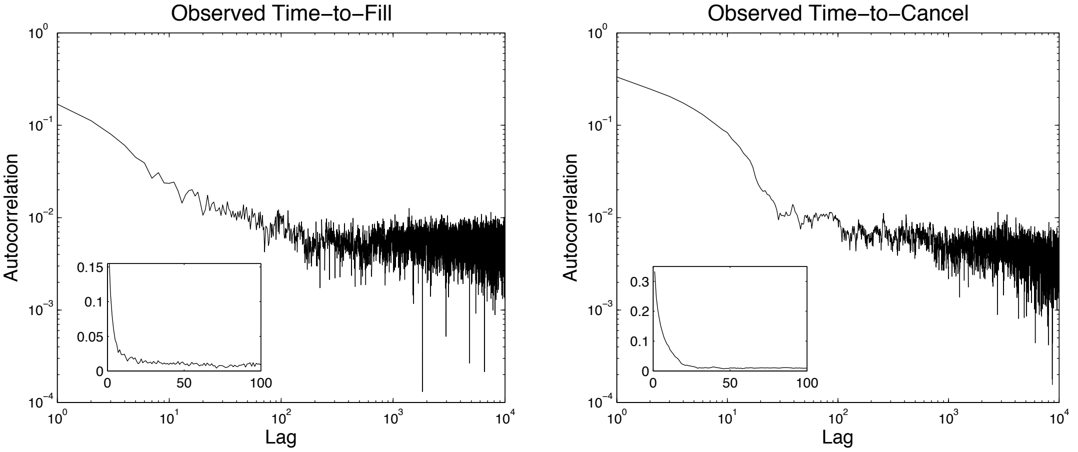


Figure 4-12: Autocorrelation of removal times for the stock AZN. For both observed time-to-fill (left) and observed time-to-cancel (right), the autocorrelation function decays slowly. The removal times are ordered by the time of removal.

Modeling Parameter Change with Cubic Splines

Our goal is to find a parametric form for the conditional distributions $P_e(\tau|\cdot)$ and $P_c(\tau|\cdot)$ as a function of conditional values. In this subsection I introduce a method for modeling q-Weibull parameter change on conditional values using smoothing cubic splines.

Before applying smoothing method with cubic splines I calculate confidence intervals for the parameter estimates using the empirical bootstrap method (Efron, 1979, 1981). The method first creates multiple sets (100 sets in our analysis) of orders sampled with replacement from the empirical distribution. Each set contains the same number of orders as in the original bin. Then I model the time-to-fill and the time-to-cancel of each sampled set with our estimation method. By the central limit theorem, I assume that the sample of bootstrap parameters are normally distributed. From this I can calculate the confidence intervals for the parameter estimates using the means and variances.

One important issue with the bootstrap method that needs to be addressed is the positive autocorrelation of removal times. Figure 4-12 shows the autocorrelation of both time-to-fill and time-to-cancel for the stock AZN, ordered by the time of order removal. In both cases the observations have strong positive autocorrelations. This means the confidence intervals to be overly optimistic.

The estimated q-Weibull parameters for each bin and their confidence intervals describe conditional distribution of removal times given certain range of conditional values – i.e.

$P_e(\tau|a \leq \Delta p < b)$ where a and b are the minimum and the maximum of Δp in the bin. To model the parameter change given a fixed conditional value instead of the range, a cubic spline smoothing technique is used. A cubic spline is a set of piecewise third-order polynomials joined together with conditions that ensure continuity. A non-parametric regression using a cubic spline (Schoenberg, 1964; Reinsch, 1967) will be used to learn a function describing q-Weibull parameters given conditional values. For example a cubic smoothing spline describing the change in tail parameter q of conditional time-to-fill $P_e(\tau|\Delta p)$ is a function $g_q(\Delta p)$ that minimizes the following equation,

$$\delta \sum_{i=1}^{n} (q_i - g_q(\mu_{\Delta p_i}))^2 + (1-\delta) \int_{\min(\mu_{\Delta p_i})}^{\max(\mu_{\Delta p_i})} g_q^{''}(u)^2 du, \tag{4.17}$$

where q_i is a tail parameter estimate for i-th bin with conditional mean $\mu_{\Delta p_i}$ and δ is a smoothing parameter. By minimizing Equation 4.17 we are minimizing the residual sum of squares and the amount of curvature of the spline – to quantify the curvature, the equation integrates the square of the second derivative of the spline function g_q. The former will make the function g_q close to the data and the latter will make g_q smoother. The smoothness parameter δ determines how much weight to place on these contradictory demands. For $\delta = 0$, minimizing Equation 4.17 will result in a linear least squares fit and for $\delta = 1$ it will result in cubic spline interpolation.

Choosing a good smoothing parameter δ is an important aspect in finding a smoothing cubic spline function. Among various methods proposed in the literature (Wahba, 1975; Silverman, 1985), I used leave-one-out cross-validation to pick δ. Let's define $g_{q,\delta}^{[k]}$ be the cubic spline minimizing the following function,

$$\delta \sum_{i=1, j\neq k}^{n} \left(q_i - g_q^{[k]}(\mu_{\Delta p_i})\right)^2 + (1-\delta) \int_{\min(\mu_{\Delta p_i})}^{\max(\mu_{\Delta p_i})} g_q^{''}(u)^2 du. \tag{4.18}$$

The leave one out cross-validation function is

$$V(\delta) = \frac{1}{n} \sum_{k=1}^{n} \left(y_k - g_{q,\delta}^{[k]}(x_k)\right)^2, \tag{4.19}$$

and the cross-validation estimate of δ is the minimizer of Equation 4.19. The smoothing cubic splines for the other two q-Weibull parameters α and λ can be written similar to

Equation 4.18 and 4.19 by replacing q_i with α_i or λ_i and the spline function g_q to g_α or g_λ.

Once we find the cubic splines describing the parameter change over conditional values we can combine them with Equation 4.6 and express conditional survival function of limit order removal time model. For example the survival function for time-to-fill conditioned on Δp is

$$S^e_{qw}(\tau|\Delta p) = \left[1 - (1 - g_{q,\delta}(\Delta p))\left(\frac{\tau}{g_{\lambda,\delta}(\Delta p)}\right)^{g_{\alpha,\delta}(\Delta p)}\right]^{\frac{2-g_{q,\delta}(\Delta p)}{1-g_{q,\delta}(\Delta p)}}. \tag{4.20}$$

The conditional survival function for time-to-cancel or with the other conditional variable s can be similarly derived, which is similar to Equation 4.20.

In summary the estimation method first utilizes the survival analysis and the q-Weibull parameter estimation method described in Section 4.3 to exploit censored observations and effectively fit the parameters for the q-Weibull distributions. Then using the empirical bootstrap method, I compute the bootstrap estimates for the confidence intervals of the parameter estimates. Lastly, I model the parameter change given conditional variables s and Δp using cubic smoothing splines. In the end, the final product of this modeling method is a conditional survival function for time-to-fill and time-to-cancel, $S^e_{qw}(\tau|s)$, $S^e_{qw}(\tau|\Delta p)$, $S^c_{qw}(\tau|s)$, and $S^c_{qw}(\tau|\Delta p)$, for each stock.

Figures 4-13 and 4-14 show the change in q-Weibull parameters of conditional survival functions for time-to-fill and time-to-cancel of stock AZN respectively. One of the important difference among the models is change in scale parameter λ. When comparing two different survival functions $S^1_{qw}(\tau)$ and $S^2_{qw}(\tau)$, the overall scale of τ's sampled from them will be determined by their scale parameters λ_1 and λ_2. When the range of λ_1 is much greater than the range of λ_2, the samples from $S^1_{qw}(\tau)$ will generally have much greater variance than the samples from $S^2_{qw}(\tau)$. If the scale parameters are similar ($\lambda_1 \approx \lambda_2$), then it becomes meaningful to compare their shape and tail parameters α's and q's to study the differences of two models. Therefore, to see how the conditional variables s and Δp affect the limit order removal times, I first compare the range of scale parameters.

Figure 4-13 (a) and (b) shows change in scale parameters for time-to-fill models conditioning on s and Δp respectively for the stock AZN. For both conditional variables, the scale parameters increase as conditional increases. This indicates when the limit order is an large order or it is placed far away from the price currently trading at, it will take longer time to get executed, which is intuitively correct. What is more interesting is the amount of increase in scale parameters when the conditional variables increase. To observe the difference, the figure shows the changes in λ's in same scale in y-axis. For the stock AZN, it shows that the Δp contributes to much faster increase in λ. This indicates that the major determinant of time-to-fill of limit orders for the stock is not the size of the order, but the price. This effect may be explained by market participants' behavior. When placing limit orders, they place orders with s within the range that will make a small difference in execution times by chopping the orders into smaller pieces. As a result, unlike Δp, we don't see a significant increase in time-to-fill since we only have data set with limit orders within such range.

A similar comparison is performed for time-to-cancel figure 4-14 (a) and (b). Here we observe an opposite effect – the major determinant of time-to-cancel is not the price, but the size of an order. Also, it is worth noting that the behavior is more complex than it was for time-to-fill.

In summary, I identified that execution and cancellation times are mainly determined by the price and the size of an order respectively. In the next section, I will verify this result by simulating τ-generation model and showing that it can generate observed removal times similar to the empirical ones.

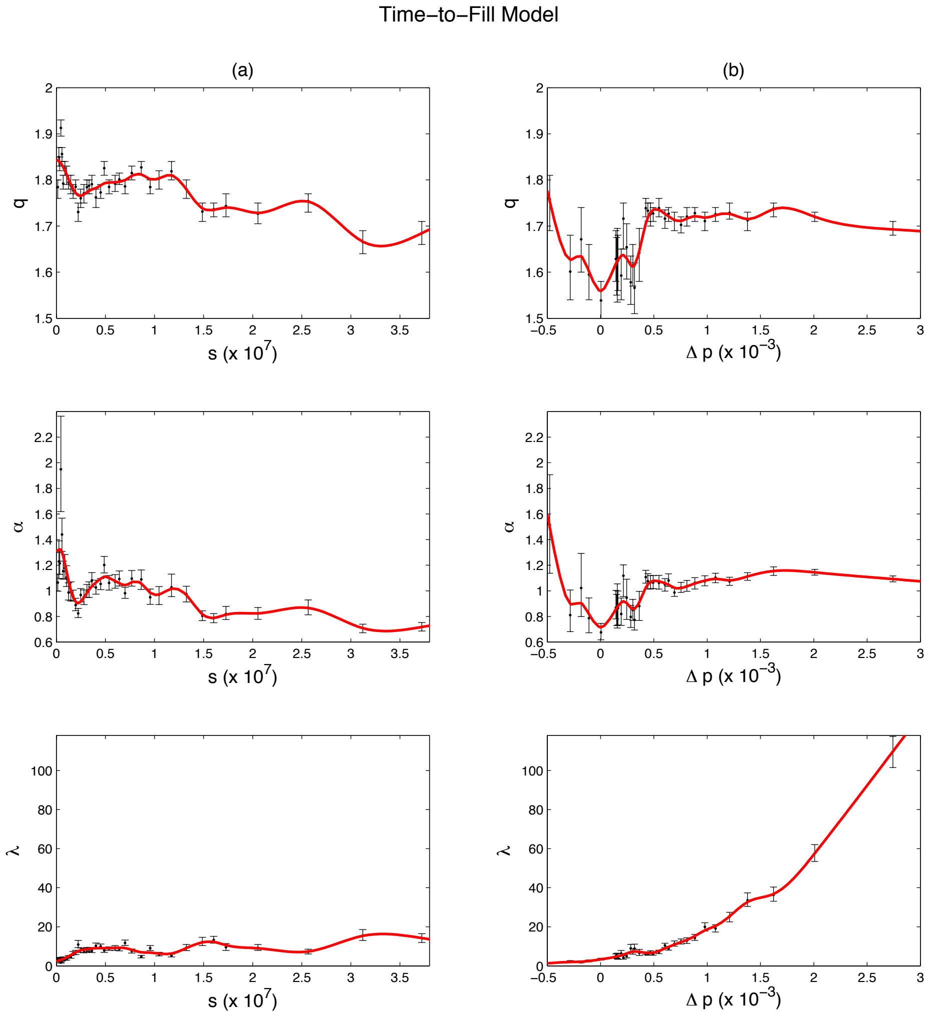


Figure 4-13: q-Weibull parameter estimates for the conditional time-to-fill model for the stock AZN. The 95% confidence intervals are calculated using the empirical bootstrap method. Cubic smoothing splines are used to model the parameter change over s and Δp. (a) $P_e(\tau|s)$: time-to-fill conditioning on s. (b) $P_e(\tau|\Delta p)$: time-to-fill conditioning on Δp.

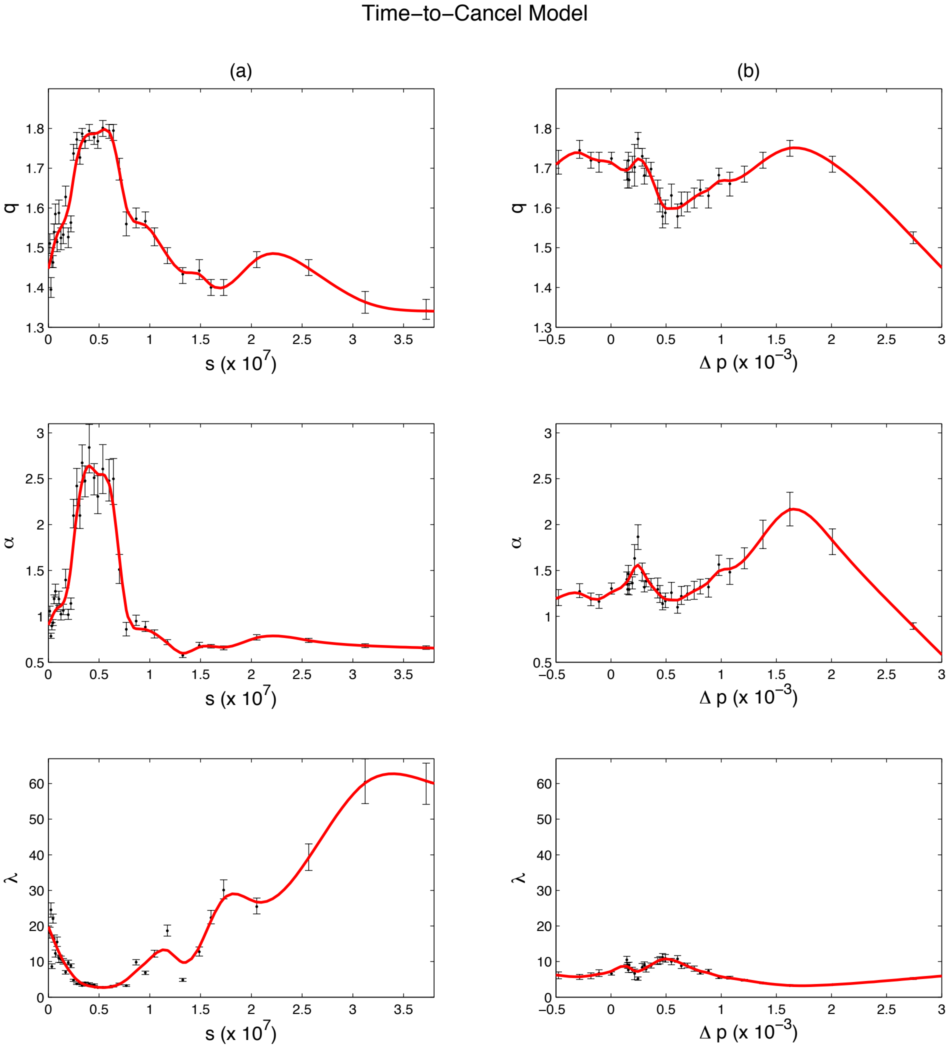


Figure 4-14: q-Weibull parameter estimates for the conditional time-to-cancel model for the stock AZN. The 95% confidence intervals are calculated using the empirical bootstrap method. The cubic smoothing splines are used to model the parameter change over s and Δp. (a) $P_c(\tau|s)$: time-to-cancel conditioning on s. (b) $P_c(\tau|\Delta p)$: time-to-cancel conditioning on Δp.

Simulation of τ-Generation with Conditional Model

In order to check if the choice of our model is valid we need to run the test with real data. Since we cannot get complete information about time-to-fill and time-to-cancel from the data set I tested how closely the model can simulate the *observed* time-to-fills and *observed* time-to-cancels. To test this N orders are simulated, where N is the number of orders in the data set. For each order i I used its empirical conditional value Δp_i and s_i to compute the conditional survival function of time-to-fill and time-to-cancel from our model (Equation 4.20). Since I discarded very large Δp's and s's during model estimation, I discarded orders with conditional values falling beyond the range of our smoothing splines functions. Then, similar to the method I used to generate the synthetic data earlier, I sampled τ_i^e and τ_i^c from estimated conditional time-to-fill and time-to-cancel survival functions and set the execution parameter $I_e(i)$ accordingly.

Figures 4-15 and 4-16 show quantile-quantile plots of real versus simulated order removal times for the stock AZN. There are four possible ways of simulating the removal times using different combinations of conditional variables: (a) $S_{qw}^e(\tau|\Delta p)$ and $S_{qw}^c(\tau|s)$, (b) $S_{qw}^e(\tau|\Delta p)$ and $S_{qw}^c(\tau|\Delta p)$, (c) $S_{qw}^e(\tau|s)$ and $S_{qw}^c(\tau|s)$, (d) $S_{qw}^e(\tau|s)$ and $S_{qw}^c(\tau|\Delta p)$. Each plot in the figure shows the quantile-quantile plot with simulated orders generated from these possible combinations.

The figures confirm the findings from the previous analysis that conditional variables Δp and s are the major determinants of time-to-fill and time-to-cancel respectively. The figure 4-15 and 4-16 (a)'s are the quantile-quantile plots of observed time-to-fill and time-to-cancel where the conditional variable Δp and s are used respectively to derive the survival functions. Among the results, they follow most closely to the 45 degree line, which indicate the closest fit to the real data. The next closest one is when Δp's are used for both execution and cancellation times (figure 4-15 & 4-16 (b)). Although the goodness-of-fit for observed time-to-cancel is almost indistinguishable from the previous one, it results less good fit for the observed time-to-fill. The other two cases where execution times are generated using the order size s result much worse fits, which clearly indicates s is not an important factor determining the time-to-fill.

Figures 4-17 and 4-18 show the survival plots of actual and simulated observed time-to-fill and time-to-cancel for rest of the 8 stocks. Here I used Δp and s to simulate execution and cancellation time respectively. This result verifies that our choice of using Δp and s to formulate conditional time-to-fill and time-to-cancel models respectively is the plausible one. All models observed removal times close to the actual distributions.

In the remaining part of this section I will discuss the economic implications of the change in the tail exponent of the conditional model. Also I will compare our modeling method to the previously known method for modeling time-to-fill using first passage time instead of the survival analysis.

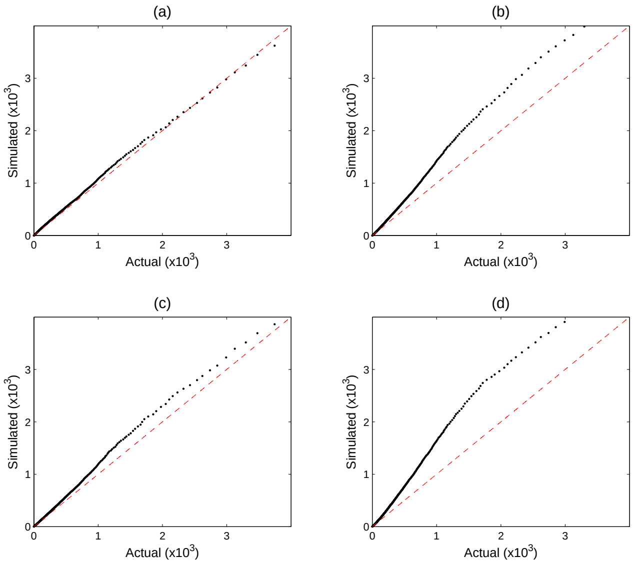


Figure 4-15: Quantile-Quantile plots for observed time-to-fill for stock AZN. In generating observed time-to-fills, all four possible combinations of conditional variables are used in τ-generation model: (a) Δp for time-to-fill and s for time-to-cancel, (b) Δp for both time-to-fill and time-to-cancel, (c) s for both time-to-fill and time-to-cancel, (d) s for time-to-fill and Δp for time-to-cancel.

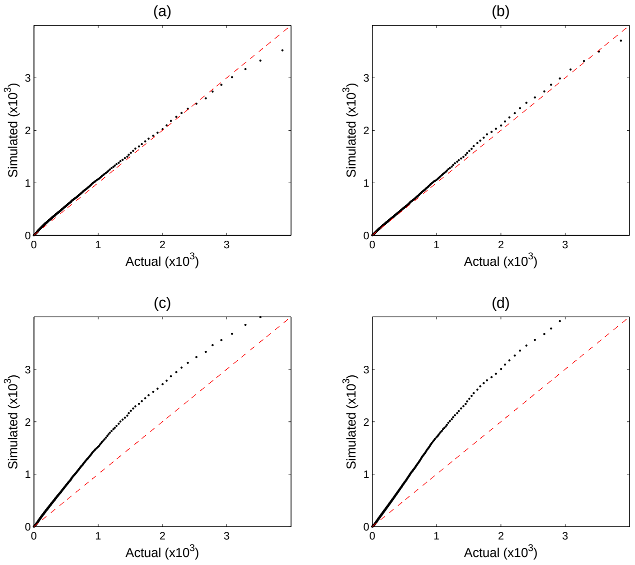


Figure 4-16: Quantile-Quantile plots for observed time-to-cancel for stock AZN. The same combinations of conditional variables are used as described in figure 4-15.

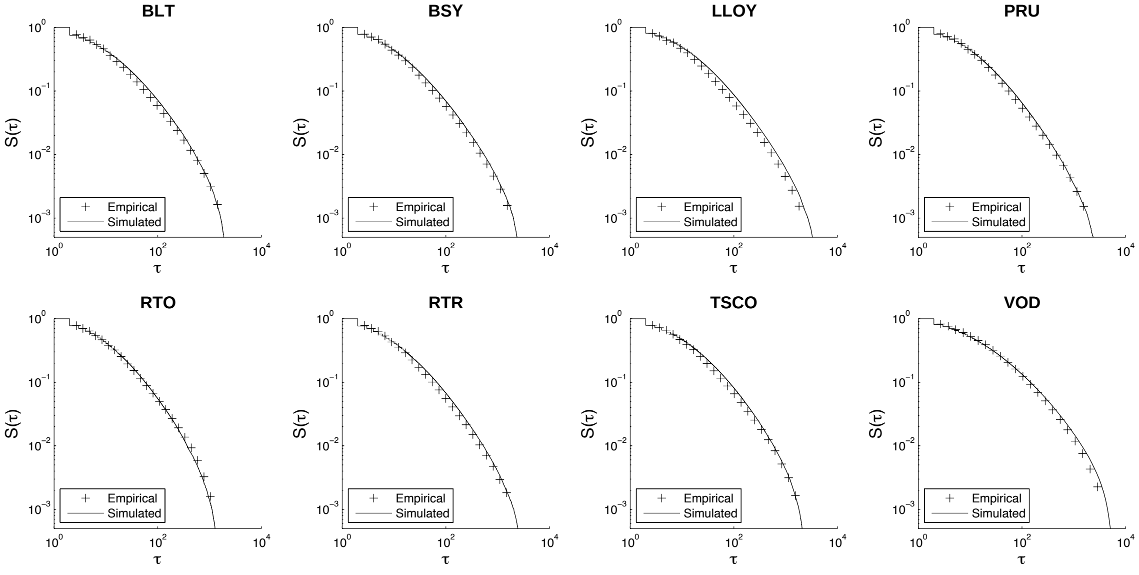


Figure 4-17: Survival plots of actual vs. simulated observed time-to-fills of 8 stocks from conditional model using Δp and s as conditional variables for time-to-fill and time-to-cancel respectively.

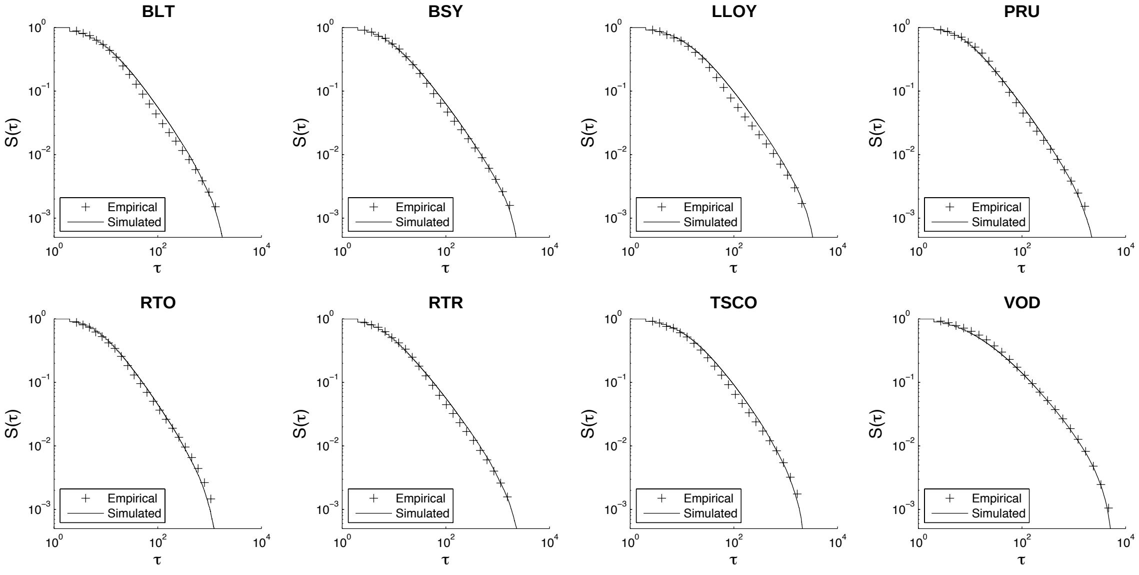


Figure 4-18: Survival plots of actual vs. simulated observed time-to-cancels of 8 stocks from conditional model using Δp and s as conditional variables for time-to-fill and time-to-cancel respectively.

4.4.3 Tail Exponent γ of Removal Times

When the q-Weibull distribution has a fat tail ($q > 1$) it asymptotically converges to a power law of the form $P(\geq \tau) \propto \tau^{-\gamma}$ for large τ. The tail exponent γ can be written in terms of the tail and shape parameters q and α as

$$\gamma = \alpha \times \frac{q-2}{1-q}. \tag{4.21}$$

I can combine Equation 4.21 with the smoothing splines describing the changes of q and α over conditional values. Since we know that Δp and s are mainly determining the execution times and cancellation times respectively, I replace q and α in Equation 4.21 with spline functions $g^e_{q,\delta}(\Delta p)$ and $g^e_{\alpha,\delta}(\Delta p)$ for time-to-fill, and $g^c_{q,\delta}(s)$ and $g^c_{\alpha,\delta}(s)$ for time-to-cancel. Then I can write tail exponents for time-to-fill and time-to-cancel as,

$$\gamma^e_{\Delta p} = g^e_{\alpha,\delta}(\Delta p) \times \frac{g^e_{q,\delta}(\Delta p) - 2}{1 - g^e_{q,\delta}(\Delta p)} \quad \text{and} \quad \gamma^c_s = g^c_{\alpha,\delta}(s) \times \frac{g^c_{q,\delta}(s) - 2}{1 - g^c_{q,\delta}(s)}. \tag{4.22}$$

What Equation 4.22 describes is the change in tail exponent of removal time distributions over different conditional values Δp and s.

Unlike λ, the parameters q and α are very sensitive to the exogenous factors I ignored in my analysis. For instance, the volatility of market or the imbalance of supply and demand can have a large effect on the shape and the tail of the time-to-fill distribution. Also the heterogeneity of market participants' cancellation behavior or their order placing strategies can significantly affect the shape and tail of time-to-cancel distributions. Although such outside factors can affect the tail exponents, some interesting regularities are observed

Figure 4-19 shows $\gamma^e_{\Delta p}$ as a function of Δp for time-to-fill, and γ^c_s as function of s for time-to-cancel for the stocks RIO, SDR, and AZN. The range of tail exponents of each stock varies significantly, which makes identifying patterns difficult when we plot them altogether. Therefore I chose three stocks with similar range. Such large variance in tail exponents across different stocks is also related to the positive autocorrelations of removal times shown earlier.

For all stocks $\gamma^e_{\Delta p}$ generally slopes downward as Δp increases. The smaller tail exponents indicate a fatter tail, therefore, this implies when a limit order is placed further away from the current price its time-to-fill distribution has a heavier tail. This implies a slower rate

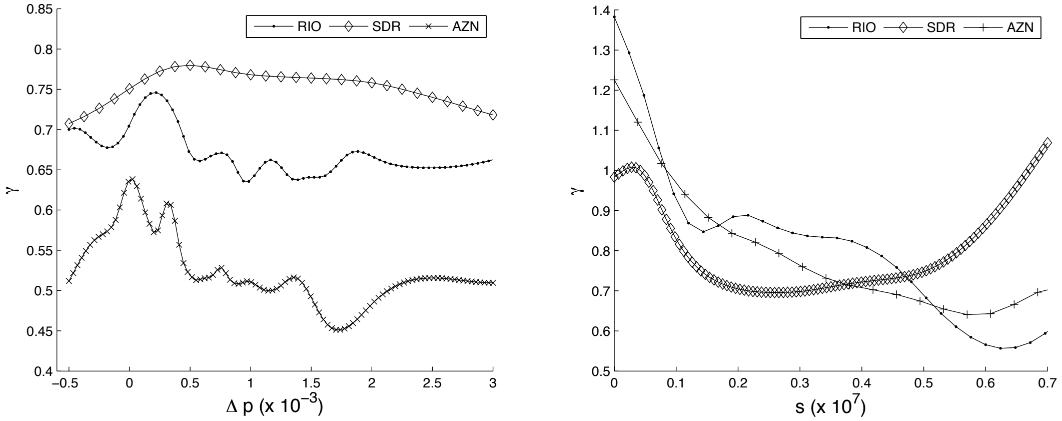


Figure 4-19: The change in the tail exponent $\gamma = \frac{(q-2)}{(1-q)} \times \alpha$ for the stocks RIO, SDR, and AZN. Tail exponents for time-to-fill conditioning on Δp's (left). Tail exponents for time-to-cancel conditioning on s's (right).

of decay of the time-to-fill survival function, meaning as the order gets placed further away from the current price the rate of increase in the probability of the order not being executed for very long period of time increases.

With some stocks for small Δp I observed an increase in $\gamma^{e}_{\Delta p}$ as Δp increases. This suggests that more aggressive orders have fatter tails, which is opposite from the previous observation. In Figure 4-19 (left) we can observe these patterns near negative regions of Δp. This may be due to chasing orders in situations when there exists a large imbalance between supply and demand in the order book. For example when there are many more buyers than sellers in the market, the price will go up. In this case the chasing orders are buy orders placed very near the trading price (small Δp). Since the price is rising, the time-to-fill will be longer than the usual case where there is no such imbalance. If more aggressive orders are placed when the imbalance is greater, such patterns can be observed.

For γ^{c}_{s} we generally observe downward slope as function of s, which implies that the time-to-cancel distribution has heavier tails for larger order sizes. The stocks RIO and AZN in figure 4-19 (right) show such patterns. This implies that market participants who place larger orders will generally have more patience. In some stocks, I observed a peak or multiple peaks creating a 'U' pattern similar to the stock SDR in figure 4-19 (left). My hypothesis is that such pattern emerges as result of the heterogeneity of market participants having different cancellation strategies or the cluster of participants with similar cancellation strategies but different size in trading capitals. Although such explanations of these patterns

Stock	$\gamma^e_{\Delta p}$ (time-to-fill)		γ^c_s (time-to-cancel)	
	5 percentile	95 percentile	5 percentile	95 percentile
AVZ	0.52	0.67	0.78	0.98
AZN	0.47	0.61	0.69	1.06
DEB	0.82	1.74	0.90	1.70
MNU	2.55	-	0.67	2.25
RIO	0.64	0.72	0.66	1.18
RNK	0.96	1.44	0.93	1.15
SAB	0.65	0.86	0.80	1.02
SCR	0.83	6.54	0.89	0.97
SDR	0.48	0.77	0.73	0.99
SGC	1.05	2.96	0.83	1.44

Table 4.3: 5 and 95 percentiles of tail exponents predicted from are removal time models. The missing value indicates that $q < 1$, which means the q-Weibull does not asymptotically converge to a power law.

are not closely verified in this study, it can be studied rigorously using individual's trading patterns from the data.

In many cases the tail exponents of removal time distributions had values smaller than 1, which indicates the first moments of removal times do not exist. Table 4.3 shows 5 and 95 percentiles of $\gamma^e_{\Delta p}$'s and γ^c_s's computed from the conditional values in the data set. The missing value indicates having $q < 1$, which does not asymptotically converge to power law.

This is significant because when we measure the removal times, we cannot describe the measure in terms of expectation. For example when the time-to-fill model estimates tail exponent less than 1, the model implies the expected time-to-fill is infinite because far too many orders have probability of not executing for significantly long period of time.

An alternative way of thinking about removal time measure is by putting an upper limit on time. For example if the upper limit is T, we can compute a probability of order being executed between 1 and T seconds and the expected time-to-fill within this range,

$$\begin{aligned} \{\text{Probability of order removal within } T \text{ seconds}\} &= 1 - S^e_{qw}(T) \qquad (4.23) \\ \{\text{Expected removal time within } T \text{ seconds}\} &= \sum_{\tau=1}^{T} f^e_{qw}(\tau) \cdot \tau. \end{aligned}$$

Figure 4-20 shows the change in expected removal times and mean of empirical observed removal times over different conditional values for the stock AZN. The expected removal times are calculated using one continuous trading period (8 1/2 hours) as the upper limit

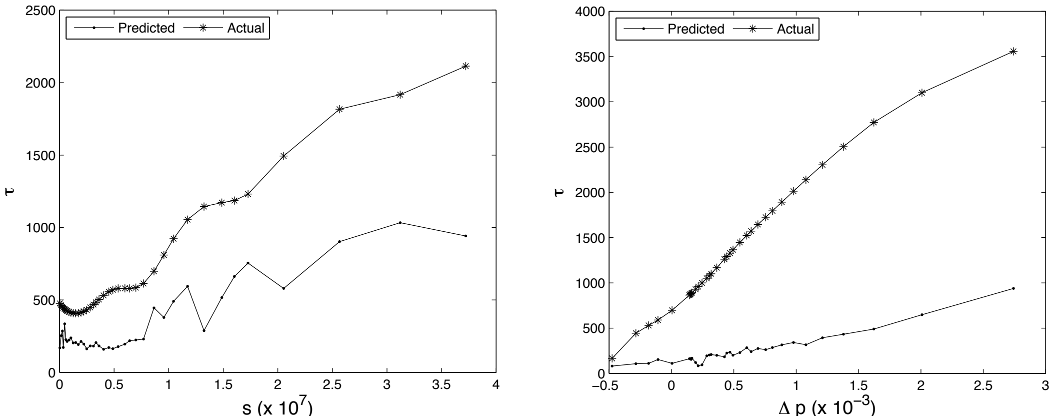


Figure 4-20: Expected and empirical removal times vs the size s and the price improvements Δp. The left plot shows time-to-cancel given s, and the right plot shows time-to-fill given Δp.

T in Equation 4.23. For both time-to-fill and time-to-cancel, our model predicts higher removal times than the observed times. Since our model predicts removal times assuming that no censoring will occur, its expectation is a prediction of true removal times rather than the observed one after censoring. The difference in the two gets larger as conditional values increase, which indicates the recovery of the problem of biased censoring – the removal times are biased towards shorter times.

Next I will explore the prospect of measuring time-to-fill indirectly via constructing hypothetical execution times from transaction data. I will use the above time-to-fill measure to compare the results.

4.4.4 Time-to-Fill Measures Using First Passage Time vs. Our Method

When modeling time-to-fill, another way of exploring the information contained in canceled orders is to construct hypothetical limit order executions using empirical data. This approach was first proposed by Handa and Schwartz (1996) and has been used by Angel (1994), Battalio et al. (1999), and others.

The first passage time is defined to be the first time the transaction price reaches or crosses the limit price. For example suppose a canceled buy order is placed at £100.00. The first time after the order placement that a transaction occurs at a price equal to or lower than £100.00 is the first passage time of that canceled order.

Lo et al. (2002) compared first passage times and actual limit orders using New York

Stock Exchange data from 1994 to 1995. From the statistical tests they revealed some serious biases, which makes the first passage time an unreliable indicator of actual execution times. The major problem with the first passage time approach is it only refers to a lower bound of execution time of a canceled order. This is because the approach ignores the placement of the canceled order in the queue. If the order is at the top of the queue, the first passage time is equal to the actual execution time, however, if the order is at the bottom of the queue, the order may not be filled with first incoming effective market order, thus, the actual execution time may be greater than the first passage time. In NYSE data, Lo et al. (2002) showed that such effect generates significant biases. The biases were stronger with illiquid stocks since the lower placement in the queue will result a longer delay in those stocks.

In this section I explore if such bias exists in LSE data. Our data is not only more recent, but also has smaller tick size, therefore the delay due to placement may be small enough to ignore. To compare our time-to-fill model with the first passage time approach I apply the following procedure: I first merge all the transaction data into one time sequence ignoring the daily boundary. The closing time between daily boundaries are ignored. The reason for doing this is many canceled orders will be thrown out if we only use a single day transaction data. Having all the transaction data stitched together into one sequence will help us to retain most of the canceled order in our analysis. Then for every canceled order I compute its first passage time by tracing the transaction prices from the actual placement time of the order. If the condition for the first passage time is not met, I treat the observation as missing.

With hypothetical time-to-fills generated from first passage time approach for canceled and observed time-to-fills for executed orders, we can fit the data with our q-Weibull time-to-fill model. In this case none of the orders are canceled, thus, the survival analysis is not used. Once we have a set of parameter estimates, we can use Equation 4.23 to calculate the time-to-fill measure within an upper time limit T.

Figure 4-21 and 4-22 show the change in time-to-fill measures over Δp of first passage time and our model for the stock AZN and DEB using upper time limit of 3,600 seconds (1 hour). These two stocks are picked to represent liquid and illiquid stocks in our data set respectively. For stock AZN in Figure 4-21 we can see both methods produce similar time-to-fill measures except for large Δp's. This is due to orders with large Δp's being thrown out by not meeting the first passage time condition.

The difference gets greater when the method is applied to an illiquid stock. For stock DEB in Figure 4-22 we see expected time-to-fill measures are lower when the first passage time approach is applied. This indicates bias of first passage time due to ignoring placement in a queue is still significant.

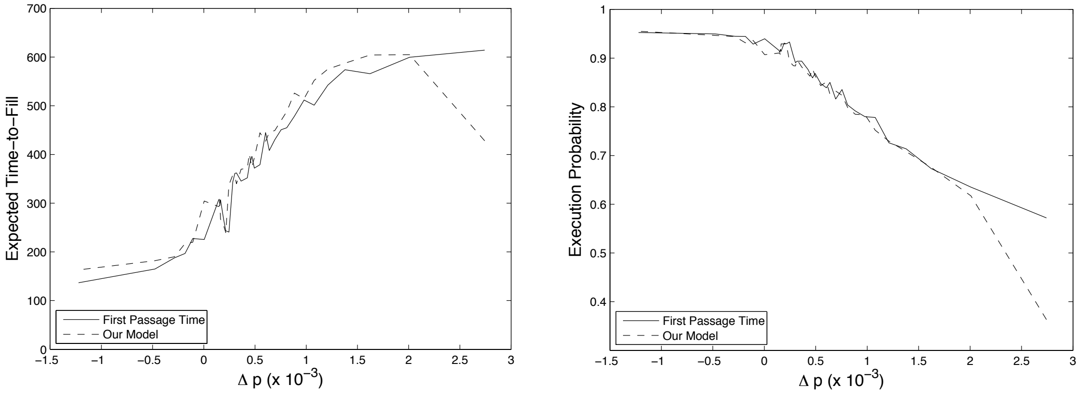


Figure 4-21: The comparison between the first passage time and our model based on survival analysis using time-to-fill measures for the stock AZN. The upper limit of 3,600 seconds is used for calculation ($T = 3600$ in Equation 4.23). The left figure shows expected time-to-fill within 3,600 seconds. The right figure shows execution probability within 3,600 seconds.

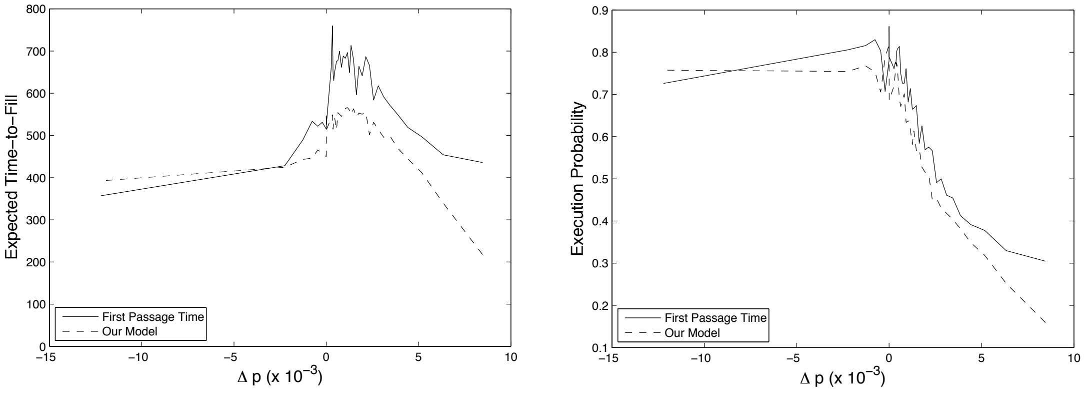


Figure 4-22: The comparison between the first passage time and our model based on survival analysis using time-to-fill measures for the stock DEB. The upper limit of 3,600 seconds is used for calculation ($T = 3600$ in Equation 4.23). The left figure shows expected time-to-fill within 3,600 seconds. The right figure shows execution probability within 3,600 seconds.

4.5 Conclusion

This chapter presents parametric models of limit order execution and cancellation times. By introducing generalized Weibull distribution, namely q-Weibull, I show the model fits the data remarkably well including removal time's diffusive behavior decaying asymptotically in time as a power law. When there are too many censored observations in the data, I find that survival analysis cannot accurately estimate the tail part of the q-Weibull distribution. Since time-to-fill process is a censoring process of time-to-cancel and vice versa, I overcome this problem by fist modeling one of the removal time processes with less number of censored observations, then deriving the other process from it. I show that resulting removal time model simulates observed time-to-fills and time-to-cancels close to the empirical distributions.

I also extend the modeling method to estimate conditional removal times introducing limit price and limit size as conditional variables. Using smoothing splines to model the change of parameter estimates over different conditional values, I present a parametric form for the conditional limit order removal time model. I find that time-to-fills and time-to-cancels are quite sensitive to limit price and limit size respectively. Given the values of these conditionals I show that conditional model simulates observed removal times closer to the empirical distributions.

I also explore the changes in tail exponents of the conditional removal time model over different conditional values. I find that as the limit order is placed at a price farther away from the midpoint price, the tail exponent of time-to-fill model becomes smaller indicating fatter tail. Similar to time-to-fill, time-to-cancel distributions generally results fatter tail when the limit size becomes larger. The fatter tail implies slower rate of decay of survival probabilities meaning we expect more un-executed and un-cancelled orders given τ for time-to-fill and time-to-cancel respectively.

Lastly I explore an alternative method of dealing with censored data problem using empirical first-passage times widely used in other literatures (Handa and Schwartz, 1996; Angel, 1994; Battalio et al., 1999). The method already has a limitation of dealing censored data problem only when modeling time-to-fills because it can only provide hypothetical time-to-fills. Similar to the result shown in Lo et al. (2002), the method performs very poorly especially when we model time-to-fills for illiquid stocks.

Chapter 5

Order Flow Simulation

5.1 Introduction

In this chapter I describe a procedure of simulating order flow using the model described in Chapter 3. I then compare the price return statistics of simulated data to the empirical data. For the comparison, I focus on three statistical properties of stock returns: fat-tailed distribution of returns, clustered volatility, and no autocorrelation of returns.

The order flow model simulates order placement and cancellation using three components: order sign, price, and cancellation time models. There is no direct link between the statistics of placement/cancellation processes and the statistics of resulting price returns. Surprisingly, the results show that the model can successfully replicate the statistical properties of stock returns observed in the empirical data.

5.2 Simulation Procedure

The order flow is simulated using order placement and cancellation models described in Chapter 3 and 4. A price of an order is sampled from IOHMM, which simulates price sequences conditioned on previous state and bid/ask spread. A side (buy or sell) of an order is sampled from FARIMA model, which simulates long memory process of transaction signs. Finally a cancellation time of an order is simulated from q-Weibull distribution estimated from the empirical data using survival analysis. As in the LSE, a simple double auction market mechanism is used to match the order and from this, transaction prices are simulated. Figure 5-1 depicts the procedure of order flow simulation.

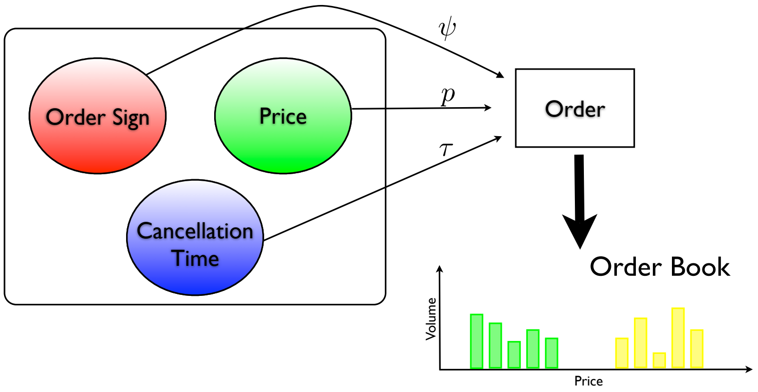


Figure 5-1: Description of Order Flow Simulation

5.3 Simulation Results

In this section, I compare the statistical properties of the empirical stock returns to the simulated returns. There has been many empirical studies of statistical properties of stock returns – fat-tailed returns distributions, clustered volatility, and market efficiency. I will address all three properties of price returns and show that our order flow model can successfully replicate them.

5.3.1 Fat-Tailed Returns

A price return at time t is a logarithm difference of mid-prices at time t and $t-1$,

$$r_t = \log(\overline{p}_t) - \log(\overline{p}_{t-1}), \tag{5.1}$$

where a mid-price at time t is an average of bid and ask prices at time t,

$$\overline{p}_t = \frac{p_{ask,t} + p_{bid,t}}{2}. \tag{5.2}$$

Figure 5-2 shows the probability densities of acutal (black) and simulated price (red) returns in semi-logarithmic scale. It shows the simulated order flow formed a fat-tailed return distribution close to the actual data. The figure also shows the probability density function of Gaussian (green), which has a parabolic shape in semi-logarithmic scale. Both the actual and the simulated returns have fatter tail than the Gaussian density.

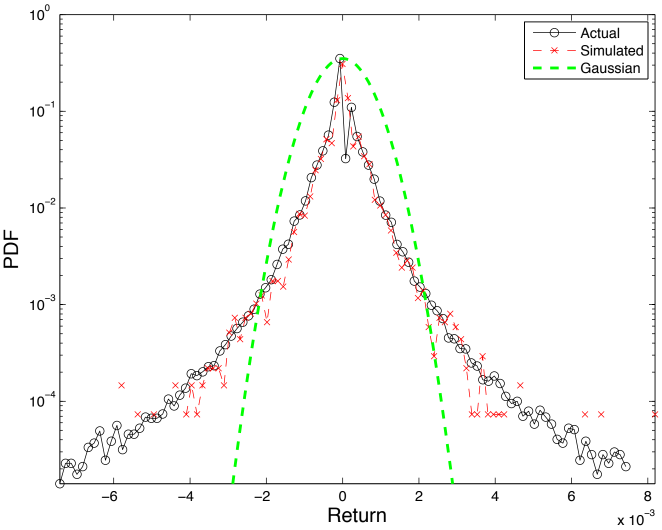


Figure 5-2: The PDFs of actual (black) and simulated (red) returns for the stock AZN. The Gaussian PDF (green) is also shown to emphasize the fat-tail of return distributions. 10^6 event times are simulated from the order flow model.

5.3.2 Clustered Volatility

Figure 5-3 shows autocorrelation of absolute return series $|r_t|$ of both actual and simulated data. Having a positive autocorrelation of absolute returns indicates a large (small) price moves are more likely to be followed by large (small) price moves. The figure shows the simulated data closely replicates the well known pheneomenon of clustered volatility.

5.3.3 Market Efficiency

The weak form of Efficient Market Hypothesis (EMH) asserts that future price returns cannot be predicted from past returns (Fama, 1965; Samuelson, 1965; Fama, 1970). This can be verified by the autocorrelation function of price returns. Figure 5-4 shows the autocorrelation of actual and simulated returns. Similar to actual data, the simulated data showed no autocorrelation, which leads to an interesting puzzle of price returns. Let's describe return r_t as the following

$$r_t = \Psi_t PI(v_t) + \eta_t, \tag{5.3}$$

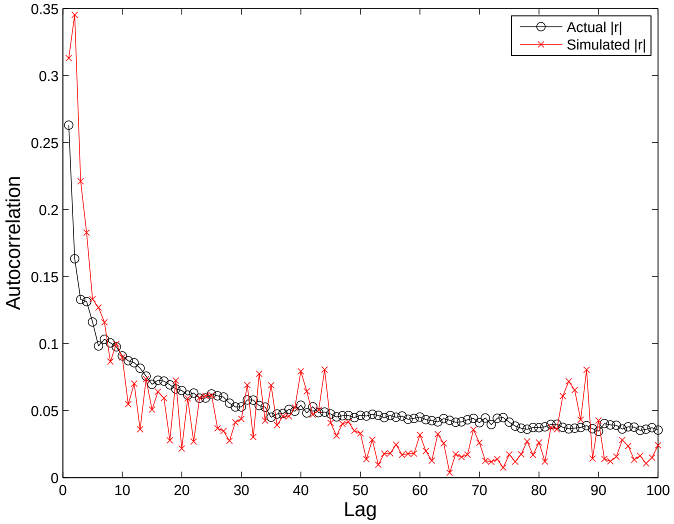


Figure 5-3: Autocorrelation of actual (black) and simulated (red) absolute return series $|r|$ for the stock AZN. 10^6 event times are simulated from the order flow model.

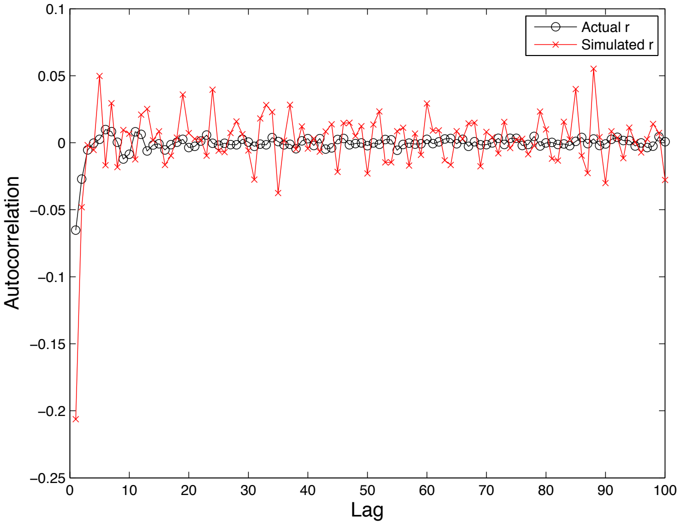


Figure 5-4: Autocorrelation of actual (black) and simulated (red) return series r for the stock AZN. 10^6 event times are simulated from the order flow model.

where Ψ_t is an transaction sign of an order flow, $PI(v_t)$ is a price impact of order size v_t, and η_t is a random noise. If Equation 5.3 is a true description of returns and assuming the variance of random noise η_t is small, return series has to be positively autocorrelated in both actual and simulated cases since we know that transaction signs are postively autocorrelated.

An alternative description of price return is introduced by Lillo and Farmer (2004),

$$r_t = \Psi_t \frac{PI(v_t)}{\zeta_t} + \eta_t, \tag{5.4}$$

where ζ_t is a measure of liquidity of the market at time t. If the transaction sign Ψ_t and the liquidity ζ_t are positively correlated, we can explain the market efficiency. In Chapter 6 I introduce a possible measure of market liquidity ζ_t to show how simulated price returns can satisfy the weak form of EMH. The further details of theory of price impact and market liquidity described in this section are described in Gerig (2007).

5.4 Conclusion

We simulated an artificial "zero intelligence" order book and showed that the model replicates statistical properties of price returns, such as clustered volatility, fat-tailed return distribution, and no predictability of future returns. We argued that the change in market liquidity explains how the order flow model satisfies the weak form of the Efficient Market Hypothesis (EMH), which asserts that future price returns cannot be predicted from past returns. This is not an obvious result since the order flow model has a predictable transaction sign process (i.e. buyer or seller initiated trades) as one of its components. In Chapter 6, we introduce a method of quantifying market liquidity from order flow data to explain this result.

Chapter 6

Liquidity Measure

6.1 Introduction

The term *liquidity* refers to different aspects of markets, such as price impact, time-to-fill, and probability of fill of various orders. Generally, liquidity means an ability to quickly trade stocks without causing a significant impact on the stock price. Although the term *liquidity* is widely used in finance literatures, its meaning is loosely defined and there is no quantitative measure for it.

In Chapter 5, I showed that our order flow model simulates price return series r_t with no autocorrelation even when the transact sign series Ψ_t is a long memory process. Such phenomenon can be explained from the following model of price return. Here I assume that random noise η_t in Equation 6.1 is small enough so that it has almost no effect on price returns. It is safe to ignore the noise since it is not clear what the noise variable is referring to in the return model. Therefore, the modified price return model is,

$$r_t = \Psi_t \frac{PI(v_t)}{\zeta_t}, \tag{6.1}$$

where $PI(v_t)$ is a price impact function of order size v_t, and ζ_t is a liquidity measure.

Our order flow model simulates long memory transact signs Ψ_t, and since all the volume v_t are fixed, we can assume that the price impact $PI(v_t)$ is constant for every transactions. In order for this model to simulate a series of returns with no autocorrelation, a liquidity measure ζ_t has to scale down the price impact $PI(v_t)$ when the transact signs are highly predictive. This means the predictability of transact sign Ψ_t and liquidity measure ζ_t are positively correlated.

In this chapter we introduce a possible quantitative measure of market liquidity ζ_t, which affects price returns to become a no memory process.

6.2 Liquidity Measure

The quantitative measure of market liquidity is derived by answering the following questions. Given a current market state:

- What is the expected cost of immediacy paid by traders who place market orders (i.e. liquidity takers)?
- In order to save this expected cost, what is the expected wait time for traders who place limit orders (i.e. liquidity providers)?

To compute the expected wait time for saving the expected cost of immediate executions, the above two questions need to be answered. Notice that the liquidity measure I am proposing is the relative measure, meaning the increase (decrease) in liquidity can occur in two ways. For instance a longer expected wait time refers to a greater market liquidity and this can occur by liquidity takers (market order placers) taking lesser liquidities (standing limit orders) by placing lesser market orders, or liquidity providers (limit order placers) providing more liquidities by placing more limit orders. In both cases, the expected wait time will increase indicating increase in liquidity. The similar argument can be made for the shorter expected wait time, which refers to decrease in market liquidity.

Our approach is to measure ζ_t using the orders placed within the fixed interval in event times. The expected cost of market orders are computed by taking an average of the absolute differences of the transacted price and the same best price of market orders. For instance when a share of buy market order is transacted at £10.10 and the same best price (bid price) at that time is at £10.00, the cost paid by the trader who submitted this market order is £0.10. I defined this to be the cost of the market order since a share could have been bought at the bid price (£10.00) by waiting. Therefore the average cost of market order $\overline{p}_m$ is defined as

$$\overline{p}_m \quad = \quad \frac{1}{N_m}\sum_{i=1}^{N_m}|p_{i,m} - p_{i,sb}|, \tag{6.2}$$

where N_m is the number of market orders submitted, $p_{i,m}$ is the i'th market order's transacted price, and $p_{i,sb}$ is the same best price when the i'th market order is submitted. Equation 6.2 provides the answer to the first question mentioned above – the expected cost paid by liquidity takers for the immediate transactions.

The expected wait time for liquidity providers to save $\overline{p}_m$ by placing limit orders can be computed using the limit order data. First let's assume that all the limit orders in our order flow data are executed. Then the average gain $\overline{p}_\ell$ by waiting is

$$\overline{p}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |p_{i,\ell} - p_{i,ob}|, \tag{6.3}$$

where N_ℓ is the number of limit orders submitted, $p_{i,\ell}$ is the i'th limit order's limit price, and $p_{i,ob}$ is the opposite best price when the i'th limit order is submitted. However, Equation 6.3 assumes the wait time τ is long enough that all the limit orders are executed. To compute the expected gain $\overline{p}'_\ell$ by waiting for a particular wait time τ, we introduce our q-Weibull time-to-fill model defined in Chapter 4 to this equation

$$\begin{aligned} \overline{p}'_\ell &= (1 - S_e(\tau)) \cdot \overline{p}_\ell \qquad (6.4) \\ &= \left[1 - \left(1 - (1-q)\left(\frac{\tau}{\lambda}\right)^{\alpha}\right)^{\frac{2-q}{1-q}}\right] \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} |p_{i,\ell} - p_{i,ob}|, \end{aligned}$$

where $S_e(\tau)$ is a survival probability, and q, α, λ are the q-Weibull parameters of the time-to-fill model. Then the expected wait time $\overline{\tau}$ for liquidity providers to save $\overline{p}_m$ can be solved by setting the Equation 6.2 equal to the Equation 6.4, and solving for τ

$$\overline{\tau} = \left[\frac{\left(1 - \frac{\overline{p}_m}{\overline{p}_\ell}\right)^{\frac{1-q}{2-q}} - 1}{q-1}\right]^{\frac{1}{\alpha}} \cdot \lambda. \tag{6.5}$$

Based on the analyses which I describe in the following sections, I found that the logarithm of $\overline{\tau}$ turns out to be a plausible measure for market liquidity

$$\zeta_t = \log(\overline{\tau}). \tag{6.6}$$

In the following section, I compute liquidity measures of simulated order flow and explain how the return series model (Equation 6.1) can produce a no memory return series.

6.3 Analyses on Simulated Data

Using simulated order flow, liquidity measures ζ_t's are computed. I used 1,000 events window of order flow data, and the window is moved 500 events at each step to compute the next ζ_t. This makes each window to contain 500 overlapping events with the prior window. I have simulated order flow of 1 million events, which results 1,999 ζ_t's. Initially I computed two different liquidity measures for each window – buy and sell liquidities. The buy (sell) liquidity is computed using only buy (sell) orders, which refers to a liquidity of demand (supply) side of the order book.

To see how liquidities change over predictability of transaction signs, I compute transaction sign probability for each transaction using FARIMA model introduced in Chapter 3. From this I separately compute averages of (1) $P(buy) - 0.5$ when $P(buy) > 0.5$, and (2) $P(sell) - 0.5$ when $P(sell) > 0.5$. The greater these values refers to the greater deviation from chance meaning greater predictability of transaction signs.

Figure 6-1 shows correlations of buy/sell liquidities and predictability of buy/sell transaction signs for the stock AZN. The high predictability of a buyer initiated transaction means there has been more buyer initiated transactions in the past. This is due to the fact that the transaction sign series are strongly and positively autocorrelated. Since there were more buyer initiated transactions in the past, we expect less liquidities in the supply side of the order book. Also more buyer initiated transactions means less seller initiated transactions, therefore, this will result more liquidities in the demand side of the order book. This explains the strong positive and negative correlations of predictability of buy transaction signs versus buy and sell liquidities in figure 6-1 (a) & (c) respectively. The similar argument can be made to explain the figure 6-1 (b) & (d) with the case when there is high predictability of seller initiated transaction. The same analysis is applied to the remaining 8 stocks and the results are persistent throughout all of them. The Pearson's linear correlation coefficient ρ's are reported in Table 6.1.

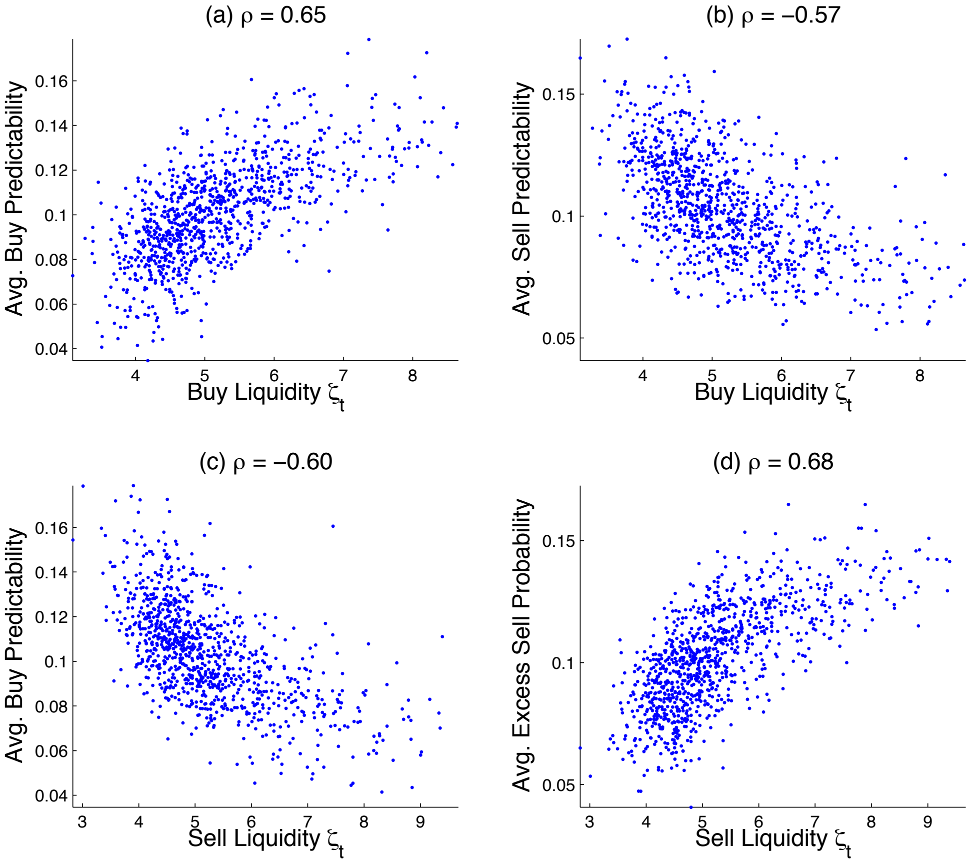


Figure 6-1: Correlations of buy/sell transaction sign predictabilities vs. buy/sell liquidities for the stock AZN. Pearson's linear correlation coefficient ρ's are reported in the figure.

Symbol	Buy Liquidity		Sell Liquidity	
	Buy Predictability	Sell Predictability	Buy Predictability	Sell Predictability
AZN	0.65	-0.57	-0.60	0.68
BLT	0.51	-0.43	-0.41	0.50
BSY	0.56	-0.45	-0.39	0.55
LLOY	0.72	-0.63	-0.63	0.72
PRU	0.68	-0.59	-0.56	0.64
RTO	0.64	-0.48	-0.34	0.50
RTR	0.66	-0.58	-0.54	0.62
TSCO	0.67	-0.51	-0.56	0.66
VOD	0.64	-0.49	-0.66	0.71

Table 6.1: Pearson's correlation coefficients of buy/sell transaction sign predictabilities vs. buy/sell liquidities for all 9 stocks.

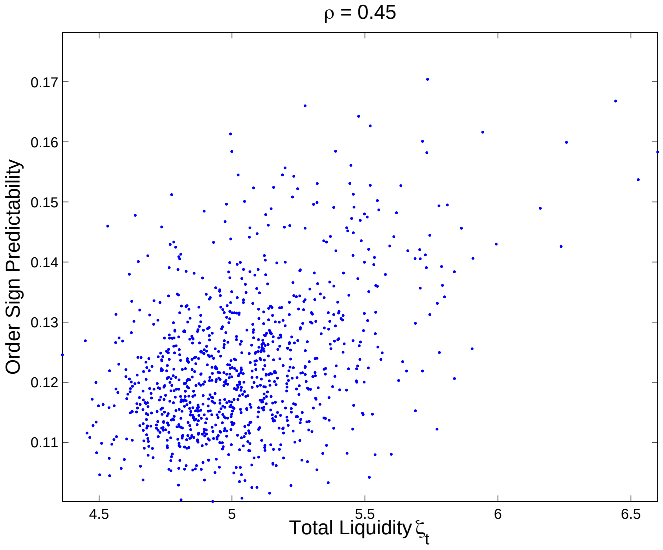


Figure 6-2: Correlation of transaction sign predictability vs. total liquidity for the stock AZN. Pearson's linear correlation coefficient ρ is reported in the figure.

Stock	AZN	BLT	BSY	LLOY	PRU	RTO	RTR	TSCO	VOD
ρ	0.45	0.28	0.33	0.27	0.23	0.43	0.37	0.53	0.42

Table 6.2: Pearson's correlation coeffiencts of transaction sign predictability vs. total liquidity for all 9 stocks.

To explain how the price return model of Equation 6.1 can generate the return series r_t with zero autocorrelation, we need to compute the total liquidity and the average transaction sign predictability of the market. The total liquidity is computed using both buy and sell orders. Similarly the average transaction sign predictability is computed by averaging the absolute value of transaction sign probabilities deviating from chance, which is $|P(buy)-0.5|$. Figure 6-2 shows a correlation of average transaction sign predictor versus the total market liquidity measure for the stock AZN. The figure shows a positive correlation between the predictability of Ψ_t and the liquidity measure ζ_t, and this explains why the proposed return model generates return series with no memory even when the transaction signs have strong positive autocorrelation. The same analysis is applied to the remaining 8 stocks and the results are persistent across all stocks.

In addition to this, I simulate price return series directly from the price return model

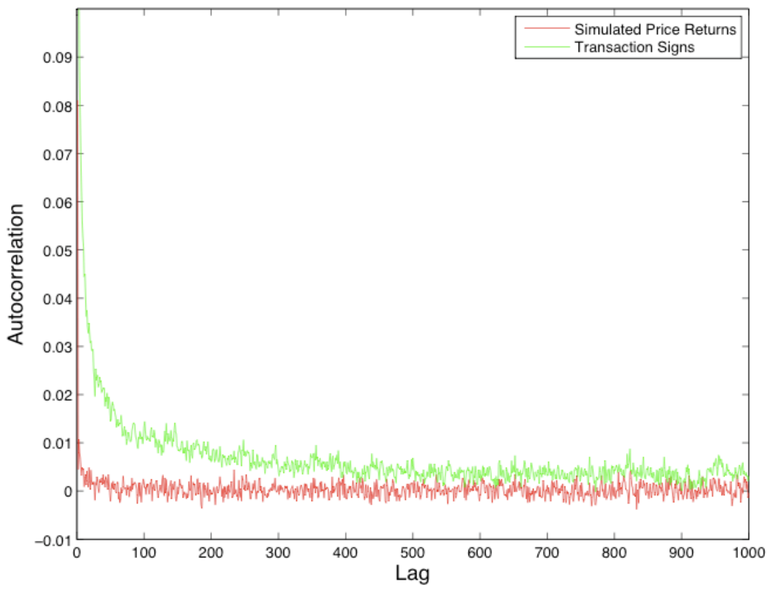


Figure 6-3: Autocorrelation of simulated transaction signs and price return series for the stock AZN

suggested in Equation 6.1. From simulated order flow, I first compute the probability of transaction sign Ψ_t and the liquidity measure ζ_t at time t, then infer the price return r_t from the price return model. Figure 6-3 shows the autocorrelations of simulated transaction signs and price returns. It shows that although the transactions signs are positively autocorrelated, the autocorrelation of price returns quickly dies out.

If transaction signs Ψ_t's have strong positive autocorrelation and are positively correlated with liquidity measures ζ_t's, the liquidity measure should also be positively autocorrelated. Figure 6-4 shows autocorrelation functions of buy, sell, and total liquidities of the stock AZN. Every other liquidity values are used to avoid the bias due to overlapping windows. In all three cases – especially the total liquidity – we see a positive autocorrelation. This result has a very interesting implication – by seeing the past liquidities we can predict the future liquidity.

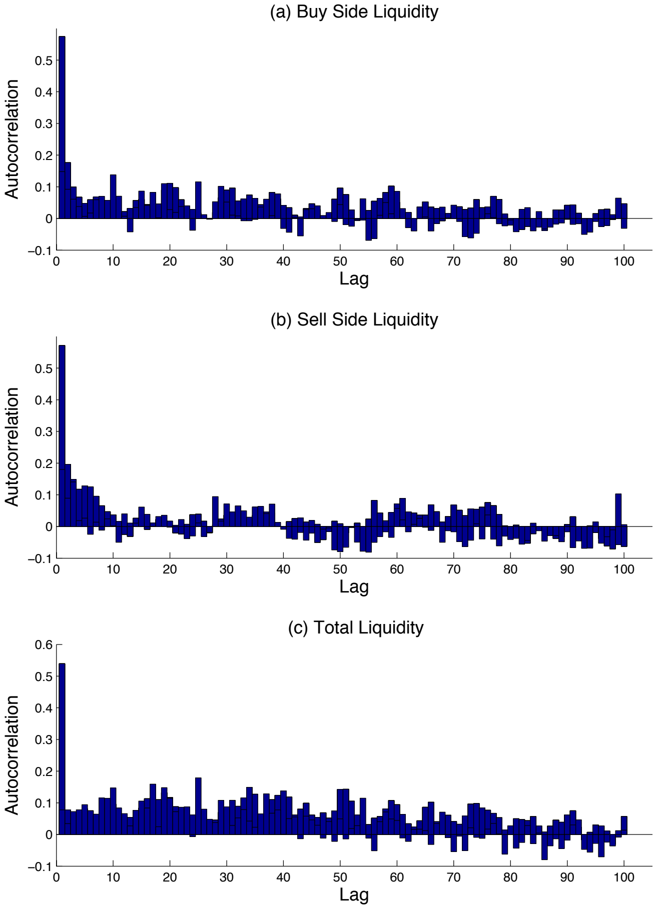


Figure 6-4: Autocorrelation functions of (a) buy, (b) sell, and (c) total liquidity measure for the stock AZN. Every other liquidity values are used to avoid the bias due to overlapping windows.

6.4 Analysis of Real Data

In this section we use our method of measuring market liquidity on a real data set. The window size of 1,000 event times is used to calculate total liquidities. Figure 6-5 shows the change in total liquidities over time for the stock AZN. The figure shows a monthly pattern in change in liquidity.

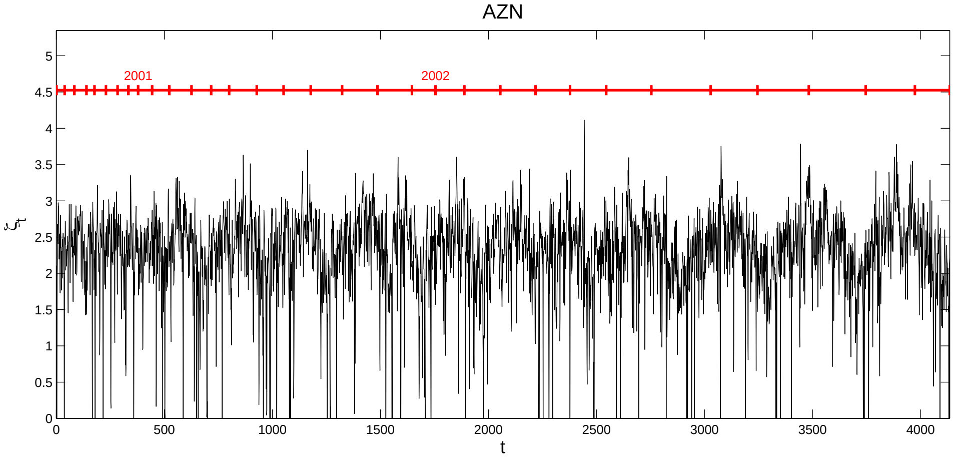


Figure 6-5: Change in total liquidities for the stock AZN.

Figure 6-6 shows change in total liquidities for the remaining 8 stocks. Similar to AZN, all 8 stocks seem to have similar patterns in change in market liquidity.

6.5 Discussion

In this chapter we introduce a method of calibrating market liquidity from the order flow data. We verify our method by showing the correlation between order sign predictability and the liquidity measure. This explains why the price return series is not autocorrelated when the transact sign series is a long memory process.

We also applied our method on real order flow data and identified patterns in change in market liquidity.

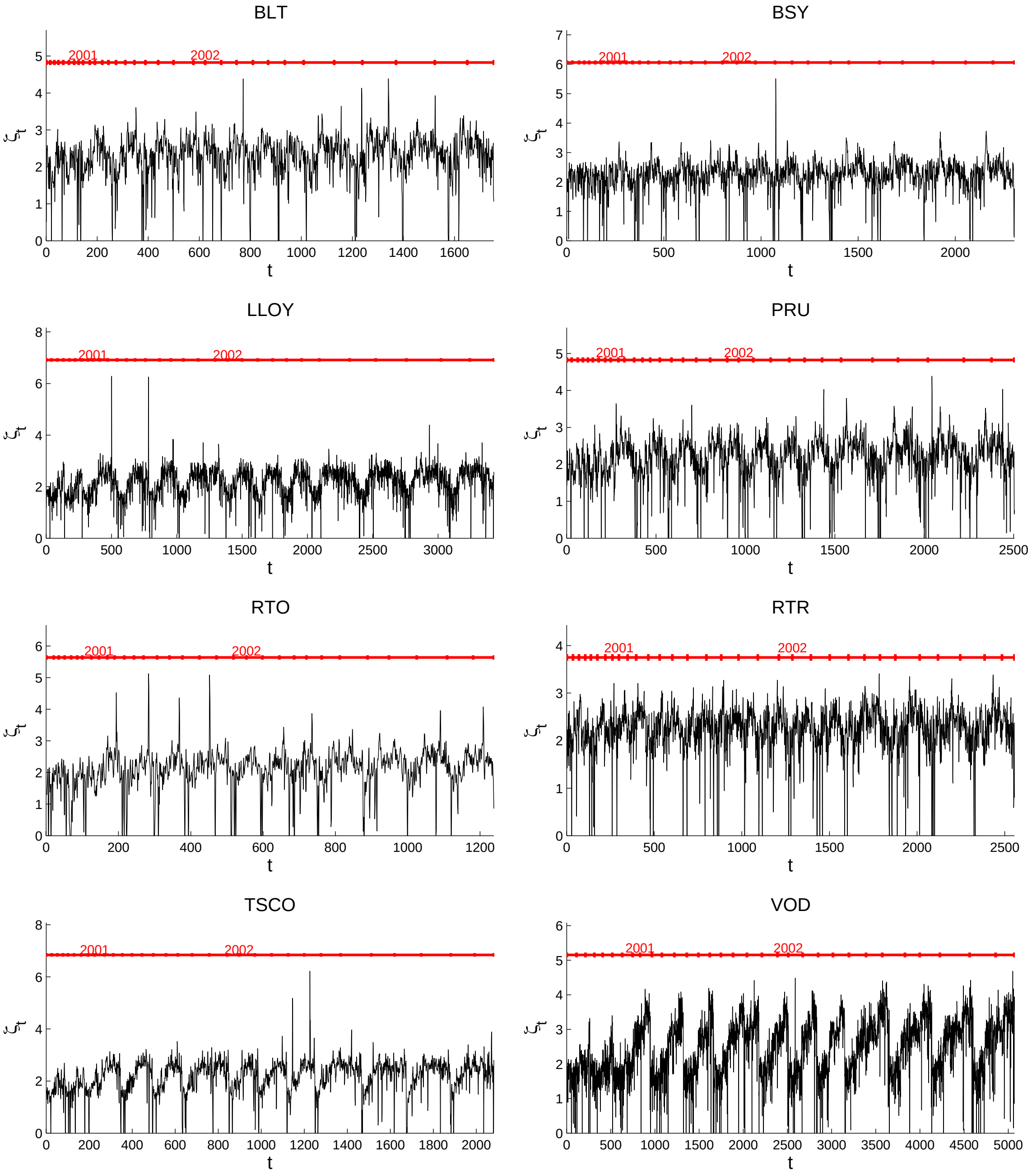


Figure 6-6: Change in total liquidities for the remaining 8 stocks.

Chapter 7

Conclusion

In the first part of this thesis, I demonstrated statistical modeling techniques to model order flow generation, which is a liquidity generation process for financial markets. I have shown that simulation of the order flow model successfully replicates various statistical properties of price returns, such as clustered volatility, fat-tailed return distribution, and no predictability of future returns.

In the second half of the thesis, I argued that the change in market liquidity explains how my order flow model satisfies the weak form of the Efficient Market Hypothesis (EMH), which asserts that future price returns cannot be predicted from past returns. This is not an obvious result since the order flow model has a predictable transaction sign process (i.e. buyer or seller initiated trades) as one of its components. Then a method of quantifying market liquidity from order flow data is introduced to explain this result.

The method of quantifying market liquidity introduced in the thesis introduces a new set of interesting future research topics in financial markets. For instance, I have shown that there exists some interesting patterns of liquidity measures in the real market data. A study of why such patterns emerges from the market data may help us better understand the market dynamics. Another interesting study is to integrate the liquidity measure into existing financial models.

Appendix A

Order Flow Model Parameter Estimates

A.1 Order Sign Model Parameter Estimates

Symbol	$\hat{d}$
AZN	0.1753
BLT	0.2012
BSY	0.1847
LLOY	0.2002
PRU	0.1800
RTO	0.2293
RTR	0.2411
TSCO	0.1933
VOD	0.2772

Table A.1: FARMA$(0, d, 0)$ parameter estimates for the 9 stocks.

An expected transact sign will be computed from FARIMA$(0, \hat{d}, 0)$ model using the following equation,

$$\hat{\Psi}_t = \sum_{j=1}^{k} \beta_{kj} \Psi_{t-j},$$

where Ψ_{t-j} is a transact sign at time $t - j$. Given a gamma function $\Gamma(\cdot)$, β_{kj} is

$$\beta_{kj} = -\binom{k}{j} \frac{\Gamma(j - \hat{d})\Gamma(k - \hat{d} - j + 1)}{\Gamma(-\hat{d})\Gamma(k - \hat{d} + 1)}.$$

Then at time t, the order sign ψ_t can be sampled according to the following probabilities,

$$P(buy) = \frac{\hat{\Psi}_t + 1}{2}, \quad P(sell) = 1 - P(buy).$$

A.2 Price Model Parameter Estimates

A price series is modeled with an IOHMM. The top table shows parameter estimates for the output probabilities. At time t, a price p_t is sampled from the following conditional probability. Given state i and bid-ask spread x_t,

$$P(p_t|s_t = i, x_t) \quad = \quad \sum_{k=1}^{2} P(m_t = k|s_t = i) P(p_t|s_t = i, m_t = k, x_t).$$

All of our price models use 2 mixture components and values of $P(m_t = k|S_t = i)$ for $k = 1, 2$ are shown in the first two columns of the top table. The rest of the table show parameters for the second conditional probability in the above equation, which is a conditional Gaussian as the following,

$$P(p_t|s_t = j, m_t = k, x_t) = \frac{1}{\sigma_{jk}\sqrt{2\pi}} \exp\left(-\frac{(p_t - B_{jk}x_t - \mu_{jk})^2}{2\sigma_{jk}^2} \right).$$

The next two tables are values for transition and initial state probabilities $P(S_t = j|S_{t-1} = i)$ and $P(S_1 = i)$. Since IOHMM states are discrete, values are presented as lookup tables.

AZN

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.047	0.953	-16.70	-83.06	0.088	0.526	0.106	0.189
2	0.656	0.344	-29.87	-30.59	3.579	0.104	17.853	0.120
3	0.508	0.492	-9.92	-21.53	-0.072	-0.048	0.058	0.070
4	0.026	0.974	-72.36	-355.42	-2.155	1.135	0.046	0.311
5	0.891	0.109	-70.10	-175.50	0.269	0.194	0.165	0.108
6	0.999	0.001	-388.65	-25.10	0.032	0.288	0.021	0.190
7	0.743	0.257	-70.17	-109.74	1.128	-0.012	0.337	0.089
8	0.888	0.112	-52.76	-7.90	0.273	-0.042	0.162	0.053
9	0.034	0.966	-193.70	-146.42	0.085	0.294	0.067	0.169
10	0.264	0.736	-171.01	-2.95	1.428	-0.060	1.674	0.050
11	0.003	0.997	-70.72	-6.40	-2.166	0.023	0.048	0.069
12	0.310	0.690	-116.18	-36.66	-0.606	-0.083	0.416	0.601

Table A.2: Parameter estimates of IOHMM output probability of order price model for stock AZN. The output p_t's are scaled by its standard deviation 0.0026.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12
1	0.161	0.009	0.000	0.007	0.001	0.054	0.003	0.598	0.007	0.001	0.159	0.000
2	0.002	0.611	0.005	0.024	0.009	0.092	0.004	0.029	0.029	0.085	0.105	0.006
3	0.000	0.001	0.793	0.000	0.000	0.203	0.000	0.000	0.002	0.000	0.000	0.000
4	0.003	0.023	0.001	0.287	0.009	0.191	0.007	0.073	0.019	0.340	0.046	0.001
5	0.008	0.044	0.003	0.041	0.186	0.471	0.000	0.012	0.096	0.032	0.106	0.000
6	0.028	0.018	0.135	0.050	0.035	0.294	0.002	0.041	0.147	0.211	0.034	0.003
7	0.001	0.002	0.000	0.007	0.000	0.004	0.969	0.002	0.000	0.015	0.000	0.000
8	0.160	0.011	0.000	0.032	0.001	0.067	0.001	0.673	0.008	0.008	0.039	0.000
9	0.010	0.012	0.005	0.030	0.011	0.522	0.000	0.008	0.387	0.010	0.004	0.000
10	0.000	0.011	0.000	0.044	0.001	0.127	0.003	0.004	0.010	0.791	0.008	0.000
11	0.010	0.013	0.000	0.011	0.006	0.040	0.000	0.015	0.001	0.007	0.897	0.000
12	0.000	0.040	0.008	0.017	0.000	0.060	0.000	0.000	0.001	0.018	0.000	0.856

Table A.3: IOHMM transition probabilities of order price model for stock AZN.

i	$P(S_1 = i)$
1	0.000
2	0.462
3	0.000
4	0.005
5	0.000
6	0.110
7	0.000
8	0.000
9	0.000
10	0.292
11	0.045
12	0.086

Table A.4: IOHMM initial state probabilities of order price model for stock AZN.

BLT

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.704	0.296	0.58	-21.35	-0.103	6.012	0.064	5.082
2	0.040	0.960	-96.25	0.82	-1.595	-0.261	0.712	0.069
3	0.974	0.026	-52.85	-111.36	0.709	-3.515	0.158	2.270
4	0.623	0.377	-97.69	-27.58	-1.310	-0.102	0.226	1.153
5	0.023	0.977	-279.92	-64.36	-0.088	0.680	1.967	0.114
6	0.170	0.830	-34.08	-106.32	0.118	1.310	1.094	16.855
7	0.041	0.959	-221.54	-120.33	-0.074	0.473	0.093	0.276
8	0.186	0.814	-194.13	-60.65	0.022	1.827	0.201	1.875
9	0.158	0.842	-197.67	-103.20	-0.082	1.116	0.127	0.632
10	0.451	0.549	-4.00	-76.44	-1.149	-0.911	0.077	0.260
11	0.501	0.499	4.37	-145.49	-0.530	-1.821	0.362	0.565
12	0.967	0.033	-1.27	-44.07	-0.110	-1.929	0.067	0.217
13	0.094	0.906	-245.64	16.20	-0.720	0.139	0.617	0.218
14	0.001	0.999	79.96	-75.23	-7.917	0.271	2.223	0.260
15	0.132	0.868	-368.75	-11.24	0.092	-0.038	0.030	0.082
16	0.393	0.607	-354.17	-88.04	0.016	0.136	0.018	0.236

Table A.5: Parameter estimates of IOHMM output probability of order price model for stock BLT. The output p_t's are scaled by its standard deviation 0.0028.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.449	0.001	0.012	0.000	0.018	0.028	0.006	0.025	0.024	0.000	0.000	0.062	0.010	0.018	0.106	0.242
2	0.001	0.687	0.000	0.009	0.000	0.096	0.000	0.008	0.000	0.118	0.011	0.000	0.000	0.000	0.000	0.071
3	0.010	0.000	0.061	0.000	0.004	0.000	0.002	0.004	0.000	0.000	0.000	0.151	0.011	0.031	0.650	0.076
4	0.000	0.064	0.000	0.755	0.000	0.012	0.000	0.020	0.000	0.039	0.008	0.000	0.000	0.000	0.000	0.101
5	0.033	0.000	0.002	0.000	0.205	0.000	0.036	0.007	0.000	0.000	0.000	0.035	0.014	0.422	0.246	0.001
6	0.050	0.226	0.000	0.006	0.000	0.603	0.000	0.002	0.002	0.017	0.003	0.001	0.000	0.000	0.000	0.089
7	0.043	0.000	0.010	0.000	0.275	0.000	0.045	0.018	0.005	0.000	0.000	0.070	0.068	0.369	0.094	0.003
8	0.038	0.000	0.000	0.000	0.002	0.000	0.002	0.254	0.020	0.000	0.000	0.306	0.008	0.008	0.063	0.298
9	0.014	0.000	0.002	0.000	0.002	0.000	0.001	0.034	0.575	0.000	0.000	0.301	0.018	0.044	0.006	0.003
10	0.001	0.445	0.000	0.015	0.000	0.045	0.000	0.002	0.001	0.356	0.005	0.000	0.000	0.000	0.000	0.128
11	0.001	0.128	0.000	0.032	0.000	0.052	0.000	0.004	0.000	0.022	0.751	0.000	0.000	0.000	0.000	0.009
12	0.006	0.000	0.023	0.000	0.002	0.000	0.002	0.081	0.069	0.000	0.000	0.721	0.003	0.001	0.005	0.087
13	0.016	0.000	0.032	0.000	0.032	0.000	0.017	0.029	0.074	0.000	0.000	0.025	0.759	0.009	0.005	0.001
14	0.002	0.000	0.014	0.000	0.176	0.000	0.017	0.002	0.009	0.000	0.000	0.002	0.003	0.768	0.007	0.001
15	0.005	0.000	0.047	0.000	0.013	0.000	0.001	0.003	0.000	0.000	0.000	0.005	0.000	0.003	0.908	0.015
16	0.014	0.006	0.016	0.002	0.000	0.005	0.000	0.043	0.003	0.001	0.000	0.026	0.000	0.000	0.011	0.872

Table A.6: IOHMM transition probabilities of order price model for stock BLT.

i	$P(S_1 = i)$
1	0.131
2	0.237
3	0.000
4	0.019
5	0.000
6	0.388
7	0.000
8	0.002
9	0.000
10	0.098
11	0.055
12	0.070
13	0.000
14	0.000
15	0.000
16	0.000

Table A.7: IOHMM initial state probabilities of order price model for stock BLT.

<u>BSY</u>

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.190	0.810	-16.75	-1.52	-2.569	-0.150	5.891	0.108
2	0.393	0.607	-152.77	-29.12	-0.639	-0.195	0.706	0.817
3	0.443	0.557	-105.00	-0.66	1.596	-0.111	0.955	0.086
4	0.089	0.911	-26.28	-17.52	-0.631	4.169	0.266	9.850
5	0.626	0.374	-5.87	-5.10	-0.060	-0.071	0.074	0.068
6	0.745	0.255	-17.58	-102.14	-1.444	-1.379	0.239	9.247
7	0.015	0.985	-322.91	-134.35	-0.556	0.531	0.835	0.255
8	0.503	0.497	-6.98	-318.00	-0.102	0.046	0.066	0.023
9	0.414	0.586	-13.53	-158.15	-0.221	-0.101	0.117	0.090
10	0.668	0.332	-61.36	-4.55	0.525	-0.023	0.259	0.118
11	0.995	0.005	-5.46	-120.61	-0.057	-1.669	0.071	0.121
12	0.287	0.713	19.69	-244.63	2.825	0.868	6.672	0.329

Table A.8: Parameter estimates of IOHMM output probability of order price model for stock BSY. The output p_t's are scaled by its standard deviation 0.0032.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12
1	0.725	0.007	0.003	0.259	0.000	0.004	0.000	0.001	0.001	0.000	0.000	0.000
2	0.009	0.886	0.002	0.025	0.000	0.002	0.000	0.061	0.011	0.000	0.000	0.004
3	0.000	0.000	0.664	0.008	0.029	0.034	0.062	0.089	0.001	0.001	0.032	0.080
4	0.182	0.012	0.035	0.573	0.016	0.000	0.013	0.102	0.013	0.002	0.048	0.003
5	0.000	0.000	0.004	0.000	0.349	0.004	0.051	0.221	0.000	0.025	0.312	0.033
6	0.001	0.001	0.680	0.000	0.000	0.076	0.005	0.211	0.000	0.000	0.000	0.025
7	0.000	0.000	0.037	0.000	0.037	0.006	0.259	0.419	0.000	0.123	0.073	0.046
8	0.000	0.001	0.037	0.003	0.016	0.006	0.139	0.580	0.011	0.075	0.039	0.093
9	0.000	0.004	0.000	0.002	0.000	0.000	0.000	0.124	0.865	0.000	0.000	0.005
10	0.000	0.000	0.005	0.000	0.040	0.004	0.275	0.255	0.000	0.260	0.138	0.023
11	0.000	0.000	0.002	0.000	0.070	0.004	0.022	0.109	0.000	0.009	0.771	0.014
12	0.000	0.000	0.050	0.003	0.028	0.003	0.031	0.566	0.000	0.041	0.072	0.205

Table A.9: IOHMM transition probabilities of order price model for stock BSY.

i	$P(S_1=i)$
1	0.476
2	0.007
3	0.104
4	0.327
5	0.000
6	0.000
7	0.007
8	0.063
9	0.000
10	0.005
11	0.000
12	0.011

Table A.10: IOHMM initial state probabilities of order price model for stock BSY.

LLOY

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.443	0.557	-156.81	-2.25	1.810	-0.140	1.689	0.080
2	0.679	0.321	54.29	-78.95	2.860	0.150	13.252	0.151
3	0.073	0.927	-105.01	130.44	-2.032	0.693	0.060	0.508
4	0.138	0.862	-13.86	-111.04	0.253	-0.019	0.611	0.066
5	0.582	0.418	-16.57	-16.89	-0.011	-0.005	0.092	0.094
6	0.037	0.963	373.31	-35.20	0.872	0.204	3.299	0.201
7	0.040	0.960	116.86	-121.47	-1.797	0.478	0.144	0.235
8	0.702	0.298	-13.60	-369.78	-0.027	-0.262	0.904	0.383
9	0.291	0.709	-28.06	-354.38	-0.076	0.066	0.068	0.026
10	0.997	0.003	-311.18	-212.55	1.092	0.353	0.338	0.147

Table A.11: Parameter estimates of IOHMM output probability of order price model for stock LLOY. The output p_t's are scaled by its standard deviation 0.0029.

i \ j	1	2	3	4	5	6	7	8	9	10
1	0.853	0.016	0.001	0.001	0.022	0.000	0.003	0.001	0.082	0.020
2	0.024	0.485	0.000	0.012	0.207	0.000	0.005	0.002	0.255	0.011
3	0.015	0.006	0.076	0.000	0.372	0.075	0.159	0.000	0.269	0.027
4	0.000	0.001	0.000	0.626	0.000	0.000	0.000	0.000	0.372	0.001
5	0.002	0.025	0.011	0.001	0.643	0.010	0.026	0.000	0.271	0.011
6	0.000	0.000	0.015	0.000	0.065	0.899	0.002	0.000	0.017	0.001
7	0.004	0.011	0.136	0.000	0.196	0.011	0.232	0.000	0.296	0.115
8	0.034	0.022	0.000	0.000	0.000	0.000	0.000	0.905	0.019	0.020
9	0.027	0.028	0.023	0.111	0.255	0.008	0.053	0.001	0.410	0.082
10	0.056	0.021	0.005	0.001	0.092	0.005	0.091	0.000	0.481	0.248

Table A.12: IOHMM transition probabilities of order price model for stock LLOY.

i	$P(S_1 = i)$
1	0.439
2	0.367
3	0.000
4	0.000
5	0.041
6	0.024
7	0.000
8	0.021
9	0.101
10	0.007

Table A.13: IOHMM initial state probabilities of order price model for stock LLOY.

PRU

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.982	0.018	-22.84	-213.79	0.060	0.317	0.155	0.249
2	0.006	0.994	-43.15	-84.95	0.473	0.659	0.291	0.315
3	0.831	0.169	-1.86	-197.77	-0.130	-0.631	0.067	0.685
4	0.987	0.013	-3.99	-66.34	-0.070	-1.596	0.079	0.223
5	0.059	0.941	-287.77	-40.53	-0.307	0.349	0.358	0.870
6	0.853	0.147	-11.95	-42.32	0.118	0.783	0.197	0.254
7	0.347	0.653	-0.37	-0.13	-0.133	-0.135	0.064	0.063
8	0.350	0.650	-178.83	-318.19	0.697	0.007	0.315	0.017
9	0.164	0.836	-68.06	-1.12	0.539	-0.084	0.557	0.061
10	0.042	0.958	-106.03	-0.93	0.255	-0.081	0.223	0.060
11	0.967	0.033	-2.59	-41.73	-0.086	-0.357	0.101	0.772
12	0.991	0.009	-18.24	-47.45	1.380	0.435	0.825	0.328
13	0.462	0.538	-6.46	-13.49	7.090	-0.056	2.248	0.117
14	0.437	0.563	-16.49	-180.28	-0.163	-0.055	0.072	0.063
15	0.447	0.553	55.54	-145.76	2.925	1.917	1.710	12.273
16	0.014	0.986	-9.87	-81.15	0.289	0.003	0.310	0.085

Table A.14: Parameter estimates of IOHMM output probability of order price model for stock PRU. The output p_t's are scaled by its standard deviation 0.0031.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0.461	0.053	0.008	0.037	0.000	0.002	0.003	0.392	0.000	0.000	0.001	0.010	0.003	0.001	0.022	0.008
2	0.451	0.054	0.030	0.178	0.000	0.008	0.006	0.172	0.000	0.000	0.009	0.021	0.004	0.001	0.045	0.020
3	0.014	0.010	0.365	0.013	0.010	0.000	0.013	0.393	0.000	0.002	0.004	0.115	0.004	0.001	0.045	0.011
4	0.026	0.005	0.011	0.745	0.001	0.000	0.005	0.149	0.000	0.000	0.008	0.030	0.002	0.000	0.009	0.008
5	0.000	0.000	0.015	0.002	0.907	0.000	0.002	0.000	0.000	0.000	0.015	0.015	0.010	0.000	0.031	0.002
6	0.003	0.001	0.000	0.004	0.000	0.962	0.000	0.016	0.001	0.000	0.000	0.000	0.000	0.000	0.008	0.004
7	0.014	0.011	0.019	0.010	0.004	0.000	0.179	0.269	0.001	0.020	0.000	0.055	0.002	0.070	0.039	0.306
8	0.262	0.021	0.054	0.117	0.000	0.002	0.018	0.403	0.000	0.012	0.006	0.035	0.005	0.008	0.026	0.029
9	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.378	0.560	0.000	0.000	0.003	0.034	0.003	0.018
10	0.000	0.000	0.002	0.001	0.000	0.000	0.007	0.012	0.274	0.692	0.000	0.001	0.001	0.002	0.005	0.003
11	0.010	0.007	0.005	0.053	0.012	0.000	0.000	0.060	0.001	0.002	0.537	0.100	0.012	0.000	0.191	0.010
12	0.076	0.019	0.067	0.254	0.000	0.002	0.035	0.125	0.000	0.002	0.109	0.237	0.004	0.001	0.047	0.021
13	0.038	0.004	0.020	0.046	0.017	0.002	0.007	0.132	0.012	0.018	0.031	0.014	0.555	0.004	0.051	0.049
14	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.066	0.014	0.003	0.000	0.001	0.000	0.897	0.002	0.006
15	0.079	0.006	0.060	0.087	0.021	0.014	0.014	0.228	0.007	0.015	0.223	0.036	0.025	0.004	0.155	0.026
16	0.002	0.018	0.004	0.001	0.000	0.000	0.035	0.540	0.003	0.002	0.002	0.024	0.002	0.015	0.011	0.341

Table A.15: IOHMM transition probabilities of order price model for stock PRU.

i	$P(S_1 = i)$
1	0.008
2	0.036
3	0.008
4	0.005
5	0.111
6	0.000
7	0.000
8	0.074
9	0.037
10	0.020
11	0.375
12	0.031
13	0.141
14	0.000
15	0.143
16	0.009

Table A.16: IOHMM initial state probabilities of order price model for stock PRU.

RTO

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.258	0.742	1.43	-188.96	-0.304	0.014	0.087	0.099
2	0.093	0.907	-132.00	-4.88	0.538	-0.085	0.390	0.067
3	0.927	0.073	34.31	-187.43	2.863	-1.222	5.916	0.678
4	0.452	0.548	-36.60	-282.89	0.013	0.030	0.140	0.020
5	0.500	0.500	-207.13	-16.32	-0.539	-0.326	0.679	0.815
6	0.741	0.259	-0.77	-72.96	-1.228	-0.088	9.945	1.382
7	0.556	0.444	-115.42	-164.43	0.924	0.373	1.205	0.237
8	0.094	0.906	-133.63	-8.86	-1.666	-0.030	0.062	0.088
9	0.223	0.777	-46.98	-0.77	0.662	-0.124	1.525	0.074
10	0.513	0.487	12.04	-108.61	0.004	0.455	0.141	0.331

Table A.17: Parameter estimates of IOHMM output probability of order price model for stock RTO. The output p_t's are scaled by its standard deviation 0.0036.

i \ j	1	2	3	4	5	6	7	8	9	10
1	0.873	0.000	0.007	0.109	0.005	0.001	0.005	0.000	0.000	0.000
2	0.000	0.856	0.006	0.137	0.000	0.000	0.000	0.000	0.000	0.001
3	0.010	0.173	0.186	0.236	0.003	0.019	0.043	0.009	0.308	0.012
4	0.012	0.134	0.022	0.669	0.001	0.007	0.123	0.001	0.029	0.003
5	0.016	0.000	0.018	0.059	0.889	0.012	0.000	0.000	0.006	0.000
6	0.006	0.002	0.068	0.149	0.006	0.317	0.001	0.001	0.448	0.000
7	0.000	0.003	0.027	0.465	0.000	0.000	0.393	0.002	0.106	0.004
8	0.000	0.000	0.016	0.023	0.001	0.007	0.007	0.943	0.000	0.003
9	0.000	0.000	0.082	0.060	0.001	0.036	0.108	0.000	0.712	0.000
10	0.000	0.015	0.006	0.014	0.000	0.000	0.000	0.001	0.000	0.963

Table A.18: IOHMM transition probabilities of order price model for stock RTO.

i	$P(S_1 = i)$
1	0.000
2	0.000
3	0.087
4	0.027
5	0.000
6	0.367
7	0.000
8	0.000
9	0.519
10	0.000

Table A.19: IOHMM initial state probabilities of order price model for stock RTO.

RTR

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.753	0.247	-63.85	13.44	-0.779	-0.324	0.259	0.234
2	0.544	0.456	-4.92	-299.91	-0.105	-0.001	0.086	0.016
3	0.323	0.677	-61.23	-0.88	1.833	-0.114	1.571	0.098
4	0.990	0.010	-55.85	-16.54	0.370	0.036	0.353	0.183
5	0.162	0.838	-148.06	-186.10	0.054	0.756	0.138	0.449
6	0.862	0.138	-1.62	-64.74	5.729	-0.181	3.323	0.545
7	0.425	0.575	-6.30	-5.67	-0.058	-0.061	0.085	0.083
8	0.997	0.003	-147.28	-73.16	1.010	-0.877	6.176	0.249
9	0.242	0.758	-203.38	-14.57	-1.062	-0.151	0.439	0.336
10	0.417	0.583	0.33	-212.17	-0.069	-0.973	0.149	0.666
11	0.602	0.398	-95.50	-35.07	-0.158	-0.230	0.114	0.092
12	0.968	0.032	-83.69	-128.62	-1.619	-1.348	0.232	2.245

Table A.20: Parameter estimates of IOHMM output probability of order price model for stock RTR. The output p_t's are scaled by its standard deviation 0.0033.

i \ j	1	2	3	4	5	6	7	8	9	10	11	12
1	0.751	0.074	0.048	0.001	0.021	0.021	0.000	0.064	0.018	0.000	0.001	0.000
2	0.001	0.621	0.069	0.100	0.151	0.007	0.034	0.008	0.000	0.002	0.003	0.004
3	0.003	0.249	0.540	0.034	0.066	0.026	0.027	0.052	0.002	0.000	0.001	0.001
4	0.000	0.354	0.017	0.434	0.116	0.004	0.062	0.004	0.000	0.005	0.001	0.004
5	0.000	0.495	0.059	0.160	0.233	0.004	0.039	0.002	0.000	0.005	0.002	0.001
6	0.017	0.372	0.181	0.058	0.040	0.209	0.066	0.032	0.018	0.000	0.007	0.000
7	0.000	0.130	0.005	0.019	0.017	0.002	0.821	0.000	0.000	0.000	0.001	0.006
8	0.010	0.081	0.420	0.016	0.014	0.023	0.000	0.425	0.010	0.000	0.000	0.000
9	0.023	0.009	0.018	0.001	0.001	0.031	0.000	0.044	0.869	0.000	0.000	0.004
10	0.000	0.125	0.054	0.181	0.573	0.000	0.016	0.001	0.001	0.048	0.001	0.000
11	0.003	0.069	0.002	0.003	0.011	0.002	0.002	0.000	0.000	0.000	0.908	0.001
12	0.000	0.475	0.247	0.044	0.017	0.000	0.194	0.000	0.000	0.001	0.001	0.020

Table A.21: IOHMM transition probabilities of order price model for stock RTR.

i	$P(S_1 = i)$
1	0.057
2	0.051
3	0.517
4	0.001
5	0.000
6	0.141
7	0.000
8	0.123
9	0.090
10	0.001
11	0.000
12	0.018

Table A.22: IOHMM initial state probabilities of order price model for stock RTR.

TSCO

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.969	0.031	-4.27	-258.91	-0.397	-2.041	0.014	0.026
2	0.832	0.168	86.23	6.34	2.177	-1.122	13.537	0.150
3	0.877	0.123	-375.59	-158.10	-0.000	0.186	0.010	1.004
4	0.883	0.117	-151.10	-51.77	-0.270	0.472	0.099	1.155
5	0.077	0.923	-344.75	-40.36	-0.524	-0.162	0.167	1.236
6	0.750	0.250	25.59	-146.15	0.839	0.126	0.136	0.493
7	0.432	0.568	-328.59	-15.90	1.108	-0.403	0.413	0.024
8	0.239	0.761	-130.83	3.45	1.764	0.423	1.988	0.020
9	0.728	0.272	-69.72	2.07	1.697	-0.321	16.046	0.148
10	0.230	0.770	-4.41	-0.05	-0.397	0.000	0.014	0.010

Table A.23: Parameter estimates of IOHMM output probability of order price model for stock TSCO. The output p_t's are scaled by its standard deviation 0.0027.

i \ j	1	2	3	4	5	6	7	8	9	10
1	0.426	0.003	0.032	0.000	0.000	0.000	0.050	0.128	0.001	0.359
2	0.037	0.204	0.177	0.001	0.000	0.015	0.171	0.050	0.004	0.341
3	0.012	0.023	0.276	0.017	0.002	0.017	0.230	0.089	0.003	0.333
4	0.001	0.002	0.247	0.744	0.000	0.000	0.003	0.000	0.000	0.003
5	0.001	0.030	0.033	0.003	0.860	0.003	0.019	0.004	0.046	0.001
6	0.000	0.019	0.071	0.000	0.000	0.505	0.076	0.066	0.000	0.263
7	0.005	0.025	0.341	0.000	0.000	0.021	0.138	0.131	0.000	0.338
8	0.067	0.023	0.159	0.000	0.000	0.010	0.117	0.163	0.001	0.460
9	0.004	0.066	0.016	0.000	0.036	0.003	0.004	0.005	0.839	0.029
10	0.066	0.023	0.212	0.000	0.000	0.006	0.089	0.111	0.000	0.494

Table A.24: IOHMM transition probabilities of order price model for stock TSCO.

i	$P(S_1 = i)$
1	0.000
2	0.000
3	0.041
4	0.000
5	0.134
6	0.000
7	0.000
8	0.038
9	0.652
10	0.135

Table A.25: IOHMM initial state probabilities of order price model for stock TSCO.

VOD

i	$\omega_{i,m}$		$B_{i,m}$		$\mu_{i,m}$		$\sigma_{i,m}$	
	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$	$m=1$	$m=2$
1	0.585	0.415	-275.98	-110.37	0.005	0.223	0.014	0.194
2	0.519	0.481	-56.68	477.13	-0.486	2.114	0.021	3.845
3	0.172	0.828	-52.52	-40.34	-0.498	-0.529	0.021	0.021
4	0.561	0.439	-38.02	-39.93	-0.093	-0.089	0.074	0.074
5	0.450	0.550	0.27	-3.84	-0.231	-0.222	0.057	0.058
6	0.527	0.473	-51.87	-49.92	-0.497	-0.503	0.021	0.021
7	0.994	0.006	-100.81	-248.50	2.701	-0.024	8.412	0.024
8	0.883	0.117	-23.13	-223.56	-0.063	1.707	0.069	3.161
9	0.038	0.962	-131.59	40.46	-0.377	0.547	0.017	0.022
10	0.047	0.953	-28.51	0.48	-1.489	-0.001	0.205	0.010
11	0.091	0.909	-77.48	142.17	-0.443	1.024	0.023	0.166
12	0.342	0.658	41.72	-32.13	0.853	-0.062	0.361	0.107
13	0.799	0.201	-146.76	-132.12	-2.937	-0.797	3.403	0.108
14	0.145	0.855	-138.39	-104.20	-0.792	-0.949	0.117	0.050

Table A.26: Parameter estimates of IOHMM output probability of order price model for stock VOD. The output p_t's are scaled by its standard deviation 0.0036.

$i \backslash j$	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.648	0.000	0.000	0.069	0.075	0.000	0.047	0.086	0.000	0.003	0.000	0.072	0.000	0.000
2	0.000	0.213	0.169	0.000	0.000	0.014	0.000	0.000	0.129	0.375	0.086	0.000	0.006	0.008
3	0.000	0.049	0.369	0.000	0.000	0.001	0.000	0.000	0.126	0.382	0.059	0.000	0.005	0.009
4	0.536	0.000	0.000	0.331	0.000	0.000	0.109	0.000	0.000	0.000	0.000	0.023	0.000	0.000
5	0.162	0.000	0.000	0.000	0.465	0.000	0.005	0.368	0.000	0.000	0.000	0.000	0.000	0.000
6	0.000	0.003	0.001	0.000	0.000	0.218	0.000	0.000	0.520	0.205	0.051	0.000	0.000	0.002
7	0.478	0.000	0.001	0.170	0.034	0.000	0.257	0.038	0.000	0.005	0.001	0.014	0.000	0.000
8	0.281	0.000	0.000	0.000	0.315	0.000	0.010	0.394	0.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.064	0.177	0.000	0.000	0.055	0.000	0.000	0.233	0.404	0.051	0.000	0.004	0.010
10	0.004	0.074	0.246	0.000	0.000	0.027	0.000	0.000	0.164	0.427	0.045	0.000	0.005	0.008
11	0.001	0.172	0.155	0.000	0.000	0.033	0.000	0.000	0.159	0.363	0.102	0.000	0.005	0.011
12	0.300	0.000	0.000	0.004	0.000	0.000	0.012	0.000	0.000	0.000	0.000	0.684	0.000	0.000
13	0.000	0.127	0.261	0.000	0.000	0.000	0.000	0.000	0.128	0.295	0.036	0.000	0.151	0.001
14	0.000	0.022	0.247	0.000	0.000	0.022	0.001	0.000	0.178	0.201	0.058	0.000	0.001	0.270

Table A.27: IOHMM transition probabilities of order price model for stock VOD.

i	$P(S_1 = i)$
1	0.306
2	0.048
3	0.045
4	0.058
5	0.001
6	0.001
7	0.382
8	0.025
9	0.036
10	0.054
11	0.009
12	0.000
13	0.022
14	0.013

Table A.28: IOHMM initial state probabilities of order price model for stock VOD.

A.3 Cancellation Time Model Parameter Estimates

A cancellation time is modeled with a q-Weibull distribution. The functional form for q-Weibull PDF is

$$f_c(\tau) \quad = \quad (2-q)\frac{\alpha}{\lambda}\left(\frac{\tau}{\lambda}\right)^{\alpha-1} e_q^{-\left(\frac{\tau}{\lambda}\right)^{\alpha}},$$

where q-exponential function e_q is defined as $e_q^{-\tau} \equiv [1-(1-q)\tau]^{1/(1-q)}]$ if $1-(1-q)\tau \geq 0$ and $e_q^{-\tau} \equiv 0$ if $1-(1-q)\tau < 0$. The table below shows q-Weibull parameter estimates for all 9 stocks. The times are measured in event times.

Stock	q	α	λ
AZN	1.43	0.7024	9.3382
BLT	1.48	0.8020	6.4889
BSY	1.43	0.7197	9.8304
LLOY	1.38	0.6461	15.1451
PRU	1.50	0.8786	8.6154
RTO	1.38	0.6637	9.5057
RTR	1.40	0.6534	8.5210
TSCO	1.32	0.5983	22.1437
VOD	1.09	0.4153	118.5829

Table A.29: q-Weibull parameter estimates of cancellation time model for the 9 stocks.

Appendix B

MIT Web Market Simulator

B.1 Introduction

In order to conduct the order flow simulations and the market experiments described in Chapter 5 and Appendix C respectively, we developed a software which replicates the mechanisms of real-world market. The MIT Web Market is a software market implemented in Java, which provides a platform for different types of electronic markets on the World Wide Web. The software is designed to be used as control laboratory for software and/or human agents based simulations and experiments.

B.2 Technical Description

In this section we first describe the design of MIT Web Market software in abstract level and then discuss how the code is modularized for object-oriented programming.

B.2.1 Abstract Level Design

An abstract level design of the system concretely defines roles of electronic market in abstract level as shown in Figure B-1. The main roles are broken down into pieces as the following: order routing, match/execution, clearance/settlement, information dissemination, market administration, and market surveillance.

When an order is received from a trader (client), the market is responsible for routing the order to an appropriate internal order book, where market-makers (or auctioneers) can match it with other order(s). When multiple securities are traded, each order needs

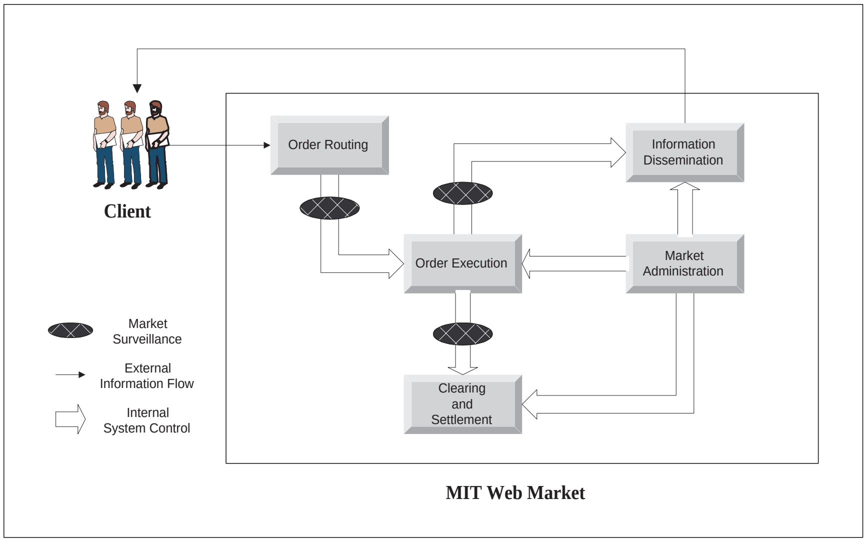


Figure B-1: An Abstract Level Design of MIT Web Market

to be routed to an appropriate market-maker (or an auctioneer) who handles the same security. After orders are matched and executed, the role of the MIT Web Market is to clear and settle the orders. The clearing/settling simply means updating the market information such as last traded price, and each trader's portfolio information. After order is matched/executed and cleared/settled, traders are notified of updated market information and their portfolio holdings, which is the role defined as information dissemination. Additionally, a market surveillance has to be taken place while orders are routed and executed. The market surveillance is to monitor the orders to make sure that they are not violating the pre-defined market rules. Lastly, the system needs to be capable of administrating the rules of the market in terms of order matching, executing, clearing, and settling.

B.2.2 Component Level Design

After the roles of MIT Web Market are defined in abstract level, the system is designed into several modules to fulfill such roles. The MIT Web Market architecture is implemented in Java 1.5 with Oracle Database *11g*, which consists of four main modules: market server, automated market-makers, database, and Graphic User Interface (GUI) for human traders. The market participants access the market via Java applet, which gets downloaded into their Web browser. The Java language provides several advantages such as portability, dynamic loading, multi-threading, and object serialization, which make it a convenient platform for

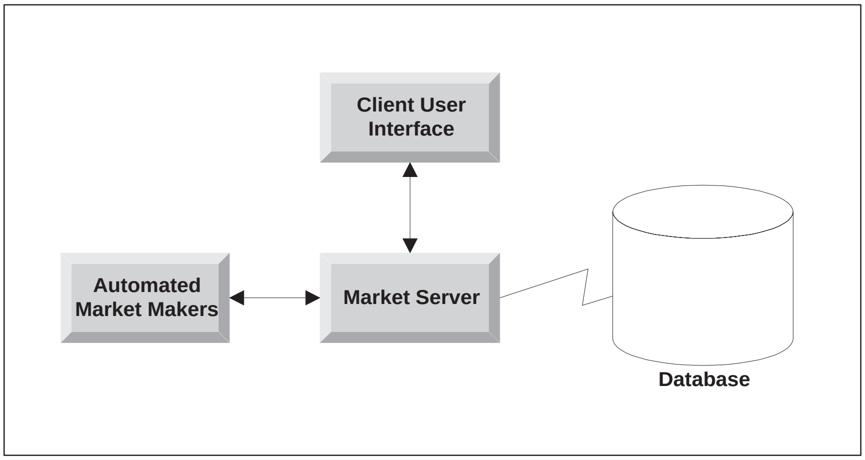


Figure B-2: A Component Level Design of MIT Web Market

implementing the inherent complexity of the market simulator. These factors are especially advantageous to our system since it uses Web technology. They made it possible for all users to interact with the system from any platforms and trade securities concurrently in the market. Also they have made possible to build a robust system with relatively limited time. The Figure B-2 represents the associations between modules. In rest of this section, we describe each module in more detail.

Market Server Module (webmarket.server package)

The market server module serves six major functions: order match/execution, information dissemination, order routing, clearance/settlement, market administration, and brokerage service.

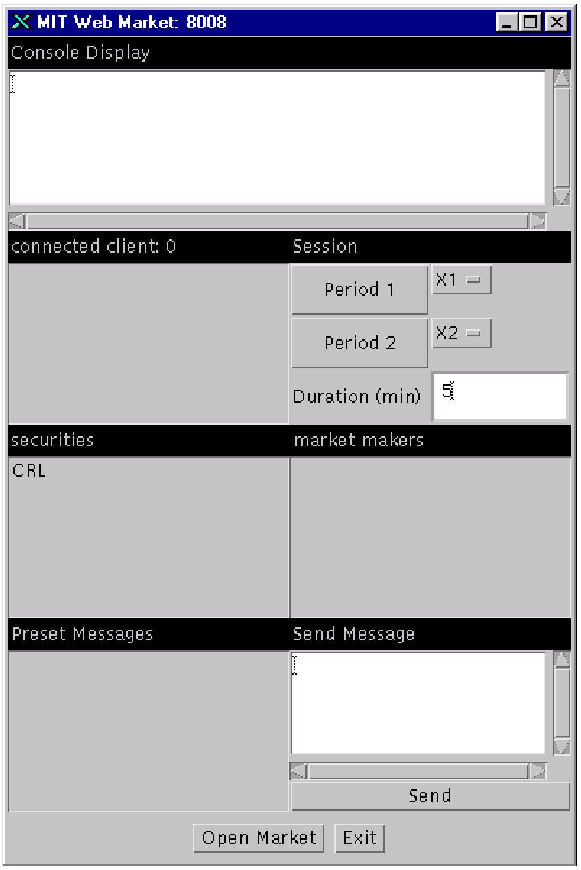
1. *Order Match/Execution* — The order match/execution plays a central role in handling and executing the orders. One of the principal advantages of object-oriented programming is that we can easily incorporate different types of order match/execution schemes into the system.

2. *Information Dissemination* — The information dissemination provides various market information to traders. The market server disseminates information in two ways; one is to disseminate information to all traders participating in the market, and the other is to disseminate information to a specific trader. In order to disseminate information to all traders, the market server uses 'Ticker Tape', which is an object storing public

information. All traders have direct access to the 'Ticker Tape'. Security information are placed on the 'Ticker Tape' when the security is traded, or when the bid/ask prices are altered by newly placed orders. The 'Ticker Tape' also provides other general market information – whether market is open or closed, opening time of market, duration of sessions, and other market news. In order to disseminate information to a specific trader, market server directly sends information to the trader using private channel. Each client maintains the channel for direct communication with the market server. The private information such as trade confirmation and portfolio holdings are sent through this channel. The market server is designed to privately send any kind of text information using this channel.

3. *Order Routing* — The automatic order routing comes naturally as part of the system. In market server, an order execution module can be an auctioneer (auction algorithm), a specialist (monopolistic market-maker) or competitive dealers (multiple market-makers). An auctioneer is a computer algorithm that matches order according to a set of pre-determined rules. The routing mechanisms are particularly easy to build under this simple rule-based algorithm. In the other hand, specialist or competitive dealer models add complexities to order routing. The automated market-maker(s) are intelligent agent(s) providing price discovery and liquidity to the market and they are designed to maintain proprietary positions and therefore bear risks. Since automated market-makers are separate module in our system, we separately discuss this in the later section.

4. *Clearance/Settlement* — After a trade has taken place, the market server clears the transaction between two counter-parties by transferring cash and securities between buyer's and seller's accounts. The role of market server as a clearinghouse is to update trader's portfolio in the database. The server creates a JDBC connection to the database and updates information. After the update, the market server provides updated portfolio information to the owner of that portfolio. Such design also contributes to flexibility of the market structure. It can be implemented with different settlement rules and clearing procedures that may apply to different securities. For example, futures market has different settlement rules from those of the options market.



Duration (min) 5

Figure B-3: A Console Window for Market Server

5. *Market Administration* — In order to administer the market, the system provides a console window for market administrator as depicted in Figure B-3. The console window is to serve five major roles: opening and closing of the market, monitoring traders who are participating in the market, monitoring the market-makers, configuring duration of sessions, and disseminate general market information to traders.

6. *Brokerage Service* — The essence of the brokerage service is the maintenance of accounts. Traders open accounts through simple registration process. When a new trader opens up an account, the market server accesses the information stored in the database on how the initial portfolio shall be created. Based on this information, the market server assigns pre-defined amount of fictitious money and shares of securities to the newly created account.

Market-Maker Module (webmarket.emm package)

The MIT Web Market is a highly configurable market in two major aspects, the market institutional structures (trading mechanisms) and the types of securities to be traded. The market structure is abstracted in the model of market-makers and auctioneers. Specifically, for each security traded in the market, different types of software market-makers can be integrated into the market in a “plug-and-play” manner.

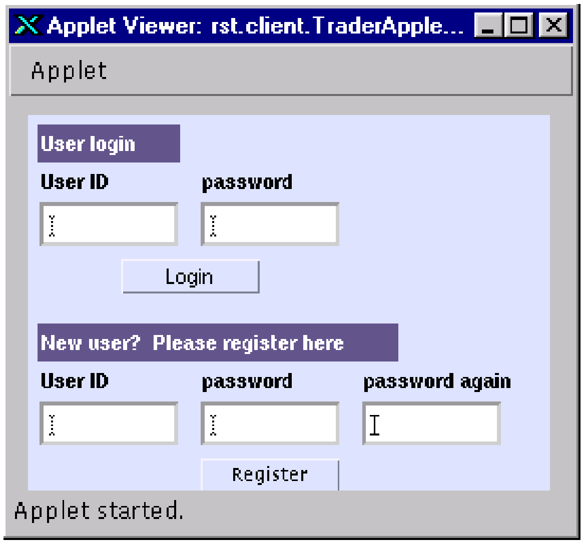
In the current MIT Web Market system, market-maker is not responsible for checking general market rules, but may wish to impose additional rules on its own clients for transactions placed through the market-maker itself. If so, market-maker should give appropriate error message of the rules that are violated to the trader. If no rules are violated, and the order is actually placed, then the market-maker must note the order to its back-end office. In addition, if this order allows some sales to proceed, the market-maker may note the sale on its back-end office to consummate the sales. Each market-maker should strive to fulfill orders eventually, of course; however, there is no requirement that it be done immediately. Also, multiple market-makers can handle the same security.

GUI Module for Human Traders (webmarket.client package)

The human traders can access the market using GUI interface developed in Java applet. The applet consists of two parts: sign up/login, and trading interface. In sign up/login, trader can either register herself as a new user or login to the market with pre-existing user ID and password. The challenge/response scheme is used to avoid transmitting the password over the network. The login window is depicted in Figure B-4.

After logging in, the client side protocol allows trader to do the followings using trading interface: place and cancel orders, obtain updates on ongoing market activity, and obtain information on its own portfolio.

1. *Place and Cancel Orders* — Orders are placed by creating an “Order” object and sending it to the market server. While orders are not executed, they can be canceled by the client. The client applet keeps track of order IDs that are previously placed, which are assigned by the market server. The client protocol sends order cancellation request along with order ID and market server uses the ID to identify the order for cancellation. If the order is not executed, the order gets canceled.



Applet started.

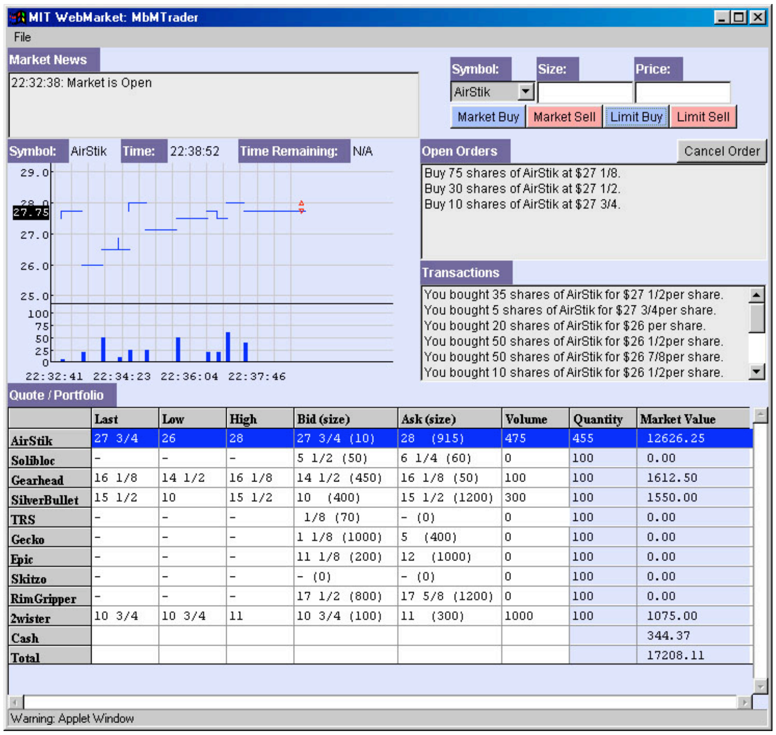
Figure B-4: A Sign up/Login Window for Trader

2. *Obtain Updates on Ongoing Market Activity* — The updates are obtained in one of two ways. The updates on market activity may be read out sequentially, or client may request to update the information from the server. The server periodically informs the ongoing sales, or other events as they have been declared of interest to a particular client.

3. *Obtain Information on Portfolio* — The number of shares held in given security are obtained by sending a holding information request to the server. The client needs to supply the security ID to receive the information. The client can also request information of amount of cash held in her portfolio in similar way.

As shown in Figure B-5, the trading interface consists of five panels: market news, trade, chart, order status, and quote/portfolio information. The information displayed in this window are obtained in two ways; 1. public information such as quote information or general market information are obtained from 'Ticker Tape' object where the market server posts public market information. 2. private information such as trader's portfolio are obtained from private channel between the market server and the trader.

1. *Market News Panel* — The purpose of the market news panel is to disseminate general market information. Information such as market state (whether market is open or closed), duration of the market, and etc.



22:32:38: Market is Open

Buy 75 shares of AirStik at $27 1/8.
Buy 30 shares of AirStik at $27 1/2.
Buy 10 shares of AirStik at $27 3/4.

You bought 35 shares of AirStik for $27 1/2per share.
You bought 5 shares of AirStik for $27 3/4per share.
You bought 20 shares of AirStik for $26 per share.
You bought 50 shares of AirStik for $26 1/2per share.
You bought 50 shares of AirStik for $26 7/8per share.
You bought 10 shares of AirStik for $26 1/2per share.

	Last	Low	High	Bid (size)	Ask (size)	Volume	Quantity	Market Value
AirStik	27 3/4	26	28	27 3/4 (10)	28 (915)	475	455	12626.25
Solibloc	-	-	-	5 1/2 (50)	6 1/4 (60)	0	100	0.00
Gearhead	16 1/8	14 1/2	16 1/8	14 1/2 (450)	16 1/8 (50)	100	100	1612.50
SilverBullet	15 1/2	10	15 1/2	10 (400)	15 1/2 (1200)	300	100	1550.00
TRS	-	-	-	1/8 (70)	- (0)	0	100	0.00
Gecko	-	-	-	1 1/8 (1000)	5 (400)	0	100	0.00
Epic	-	-	-	11 1/8 (200)	12 (1000)	0	100	0.00
Skitzo	-	-	-	- (0)	- (0)	0	100	0.00
RimGripper	-	-	-	17 1/2 (800)	17 5/8 (1200)	0	100	0.00
2wister	10 3/4	10 3/4	11	10 3/4 (100)	11 (300)	1000	100	1075.00
Cash								344.37
Total								17208.11

Figure B-5: A Trading Interface for Trader

2. *Trade Panel* — Traders can place orders using trade panel. From the pull-down menu, trader chooses a security that she wants to buy or sell, and specify size and price of an order. There are four types of orders which can be placed: market buy, market sell, limit buy, and limit sell. When market order is placed, the price is ignored by the system. If trader places an order violating the market rules, such as placing short sell when short position is not allowed, the order is ignored by the system and the order rejection message is displayed.

3. *Chart Panel* — The chart shows the price movement of particular security chosen by the trader. The trader can choose a security from pull-down menu. The chart also shows current time which is synchronized with server side time and also time remaining if there is a pre-defined duration of the market. The bid/ask prices are shown in the graph by red triangles, and the last prices are shown in blue horizontal line. At the bottom of the graph, the traded volumes are shown in bar graph.

4. *Order Status Panel* — There are two types of order status information: open orders, and transactions. The open order panel displays list of orders that are placed by the

trader, but yet been executed. The transactions panel displays executed orders that are placed by the trader. A trader is allowed to cancel an order if the order is listed in the open orders panel. Once an order is executed and listed in transactions panel, it cannot be canceled.

5. *Quote/Portfolio Panel* — The quote/portfolio panel is to display bid/ask prices, last traded price, and total traded volume of securities. It also displays trader's portfolio information. The client applet receives security information from 'Ticker Tape' object and displays them in the quote side of the panel. It also receives trader's portfolio information through private channel, calculates the market value of the portfolio, and displays them in the portfolio side of the panel.

Database Module

One of the important advantages of the integration of sub-systems is sharing of information among them. A centralized database is the core of such system. Despite the fact that the whole system may consist of a collection of geographically dispersed servers, the MIT Web Market can be viewed as a centralized system in terms of its abstract architecture. Note that within this concept, a centralized market is capable of accommodating fragmented market structures. One scenario is that within a single market, there can be multiple market-makers dealing with the same security. The market-makers receive orders via the same order routing system and are monitored by the same surveillance units. More importantly, the centralized database stores market information, security information and trader information that are shared by different sub-systems. For instance, both the brokerage and clearing service need to access the account information of an investor after a transaction. Consider the following sequence of how a investor's trading request is served. The investor places an order via some trading software provided by the brokerage service. The order arrives at the market electronically, and is further transmitted to one of the market-makers who executes the order and reports the trade information to the market. The market in turn disseminates new trade and quote information to all traders connected to the market. Simultaneously, the clearinghouse is notified of the execution. It clears the trade and makes appropriate portfolio adjustments for the two parties involved in the trade (updating cash and security position information, for example), and notify the brokerage service for the updates. Updated

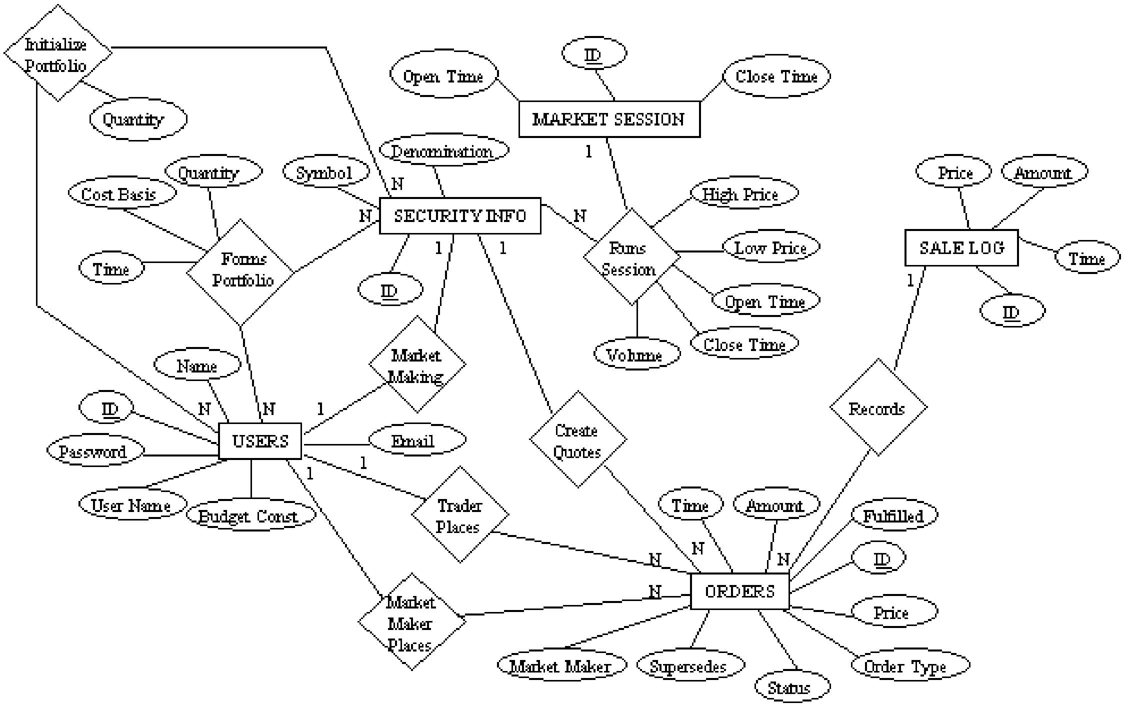


Figure B-6: An Entity Relationship Diagram (ERD) of MIT Web Market Database

portfolio information is then sent to the two investors who participate in the trade.

The MIT Web Market uses Oracle Database *11g* and it is designed based on the Entity Relationship Diagram (ERD) depicted in Figure B-6. The database is divided into two groups of tables; tables for defining the market, and tables for storing trading information.

1. *Defining Market* — There are four entities that defines the market; security information, users, initial value of portfolios, and market session. In order to setup a market, all of those information need to be provided. When trader registers as a new user, the trader's information are stored in the 'users' table. At the same time, the trader's initial portfolio is created in the 'portfolios' table based on the initial value, which are pre-defined in the database.

2. *Storing Trading Information* — There are mainly two types of trading information. One is order information and the other is sales information. The order information are created by users and they are used to route orders to execute them. After the execution, the sales information are stored in 'sale log' table. It is important to separate the two in the database since the first one is needed to show quote information,

and the second one is used to determine the price of securities.

B.3 Summary

In summary, the main features of the MIT Web Market Simulator are: automation in the trading mechanism (automated market-makers), flexible design that can accommodate various types of market structure, environment that allows interactions between electronic agents and human traders. The potential applications of the software are various financial or non-financial markets, educational or experimental markets, and tools for research in market micro-structure, which I have demonstrated in this thesis.

Appendix C

Securities Trading of Concepts (STOC)

C.0 Preface

In addition to the research relating to financial markets, the following presents my work in using market as a tool for aggregating people's preferences. The idea is to apply the market mechanism to collect consumer preferences towards product concepts through trading. This work was done in collaboration with Dr. Nicholas Tung Chan and Prof. Ely Dahan (UCLA) along with two of my research advisors Profs. Andrew W. Lo and Tomaso Poggio.

C.1 Introduction

Markets are well-known to be an efficient tool for collecting and aggregating diverse information regarding the value of commodities and assets (Hayek, 1945). They have been particularly successful in the domain of financial securities. In this chapter, we explore a novel application of the price-discovery mechanism of financial markets to marketing research using securities trading of concepts (STOC) to collect consumer preferences on product concepts. This application is motivated by the need for reliable, accurate, fast and economical means to gauge consumer preferences during new product development. It relies on the belief that markets are efficient in aggregating privately held information such as individual preferences and expectations of others preferences. It also exploits the incentive-compatible nature of markets, i.e. the fact that over- or under-valuing securities reduces the participants rewards, and the preferences of most participants for competing in games over responding to direct surveys.

In particular, we present results for multiple market experiments with live subjects in which participants express their preferences over new product concepts by trading virtual securities. The resulting securities prices are compared against preferences measured in separate studies of the same products using various stated-choice and revealed preference approaches: rank-ordered choice, the virtual concept testing (VCT) methodology developed by Dahan and Srinivasan (2000), and actual purchase decisions under controlled conditions and in actual product markets. We find that results across different market experiments in three product categories – bicycle pumps, laptop computer messenger bags and crossover vehicles – are, with one notable exception, reliable across repeated tests and predictive of stated-choice and revealed preference results. To gain a better understanding of how the STOC method may be achieving these results, we relate our market experiments with some classic examples from the experimental economics literature.

The essence of the STOC methodology centers around the establishment of virtual stock markets that trade virtual securities, each associated with an underlying product or service concept that can either exist or be at the conceptual or prototype stage. Upon entering a concept market, each participant receives an initial portfolio of cash (virtual or real) and virtual stocks. Participants are also provided with detailed information on the products (stocks) that includes specifications, images, and multimedia illustrations. A typical objective of the STOC game might be for each participant to maximize the value of his or her portfolio, evaluated at the last price prior to the closing of the market. Markets are typically open for 20 to 30 minutes. If participants play with real money, they will have the opportunity to profit from trading and will conversely bear the risk of losing money. The financial stakes in the game provide incentives for participants to reveal true preferences, process information and conduct research. If fictitious money is used, prizes can be awarded according to individuals performances. One can also reward all participants simply for their service.

As in real financial markets, stock prices are determined by supply and demand, which depend on participants evaluation of their own and others preferences for the underlying products. Thus, at the market equilibrium, prices should fully reflect all participants aggregate preference of the products. Traders make trading decisions just as they would in a financial stock market: they assess the values of the stocks, sell over-valued ones and buy undervalued ones, essentially voting on the worth of the underlying products. In this way, a stocks price becomes a convenient index of a products consumer value.

There are, of course, several well-established methods for estimating consumer preferences, e.g., surveys (Burchill and Brodie, 1997), voice-of-the-customer methods (Griffin and Hauser, 1993), conjoint analysis (Srinivasan and Shocker, 1973; Green and Wind, 1975; Green and Srinivasan, 1990), concept tests (Urban et al., 1990; Dahan and Hauser, 2002), and focus groups (Calder, 1977; Fern, 1982). However, concept markets may be a useful alternative to these methods for several reasons:

1. Accuracy: In order to win the game, participants have the incentive to trade according to the best, most up-to-date knowledge because of their financial stake in the market. STOC also captures, continuously, the ever changing pulse of the market for all participants since they can express their opinions multiple times during the course of the market rather than responding only once to a survey question.

2. Interactive Learning: A STOC market participant not only evaluates concepts on his or her own behalf, but also considers the opinions of the public at large. Furthermore, participants can observe others valuations of the virtual products and update and adjust their own valuations dynamically in the market environment. In short, learning is a crucial element in these markets.

3. Scalability: Unlike surveys, in which the number of questions asked is limited by the capacity of each respondent to answer, markets are intrinsically scalable due to the fact that each trader need only evaluate a small subset of the universe of securities. In fact, the efficiency of the market, and therefore the quality of data collected, improves with the number of participants. This extends to the number of product concepts that may be evaluated since there is no requirement that each respondent trade every security, the bounded rationality of the traders does not limit the number of concepts that can be evaluated in a STOC market.

4. Integrated Product Concepts: The STOC method is particularly useful relative to conjoint methods when a product cannot be easily quantified, delineated or represented by a set of attributes (for example, a movie script, fashion item, car body

style or piece of art). Market participants evaluate the concepts directly and market prices effectively reflect the overall viability of the concepts, including the ability of a concept to fulfill unarticulated needs. All that is required is a thorough description (and visualization) of each concept.

Of course, market-based methods for eliciting information also have certain limitations. Unlike typical marketing research techniques in which information is collected from individuals and aggregated in subsequent analysis, the market method focuses on aggregate beliefs and preferences. Individual heterogeneity is not captured well in the end, even though it enters the trading process in the form of differences in security valuation. Virtual concepts markets may be vulnerable to price manipulations and speculative bubbles because the values of virtual securities hinge on the aggregate beliefs, which are endogenously determined within the same market. Traders may form false beliefs that could cause prices to deviate from their fundamentals. And all of the behavioral critiques that have been leveled against the Efficient Markets Hypothesis in the financial economics literature apply to concepts markets as well. For these reasons, the market method must be applied with caution, and the consistency of the results must be checked through repeated STOC markets or other means of validation. The greatest level of vulnerability may occur when traders have a poor sense of their own preferences or of those of other people. This might occur, for example, when the product category is too new for traders to grasp, or when the stimuli shown prior to trading are unclear or confusing (as we demonstrate in one instance shortly). A number of practical issues arise in attempting to infer consumer preferences via the STOC method. For example:

- How many traders are needed?
- How knowledgeable does each participant need to be of the product category and concepts being studied? Of the target market?
- Do they need to be experienced at trading securities?
- What strategy do traders adopt in order to win the game?
- Are traders driven more by objectivity or personal biases?
- For how long must trading proceed in order to collect useful data?

- What, exactly, is being measured by STOC?

The present research attempts to answer many of these questions by positioning STOC in the context of prior experimental economics and prediction markets research, and by evaluating the results of empirical experiments in three product categories.

In Section C.2, we outline the market research alternatives to STOC which also measure preferences for product concepts. These include virtual concept tests of bike pumps, a simulated store selling laptop bags and stated choice surveys and longitudinal revealed preference sales data for crossover vehicles. We then summarize relevant research from the prediction markets and experimental economics literature. Section C.3 introduces the three product categories that are tested in this research, the designs of the securities representing product concepts in those three categories, and the market mechanism used to trade these securities. In Section C.4 we conjecture on how STOC works by considering alternative strategies that traders might employ, and show how stock prices capture consensus preferences. Section C.5 presents results from multiple STOC experiments, and develops a taxonomy of internal and external validity testing. We conclude in Section C.6 with a discussion of the results, possible extensions and limitations.

C.2 Background

To validate the STOC method, we compare it against alternative methods of measuring preferences for new product concepts. We also position STOC in the context of prior work on prediction markets and experimental economics.

C.2.1 Prior methods of measuring product concept preferences

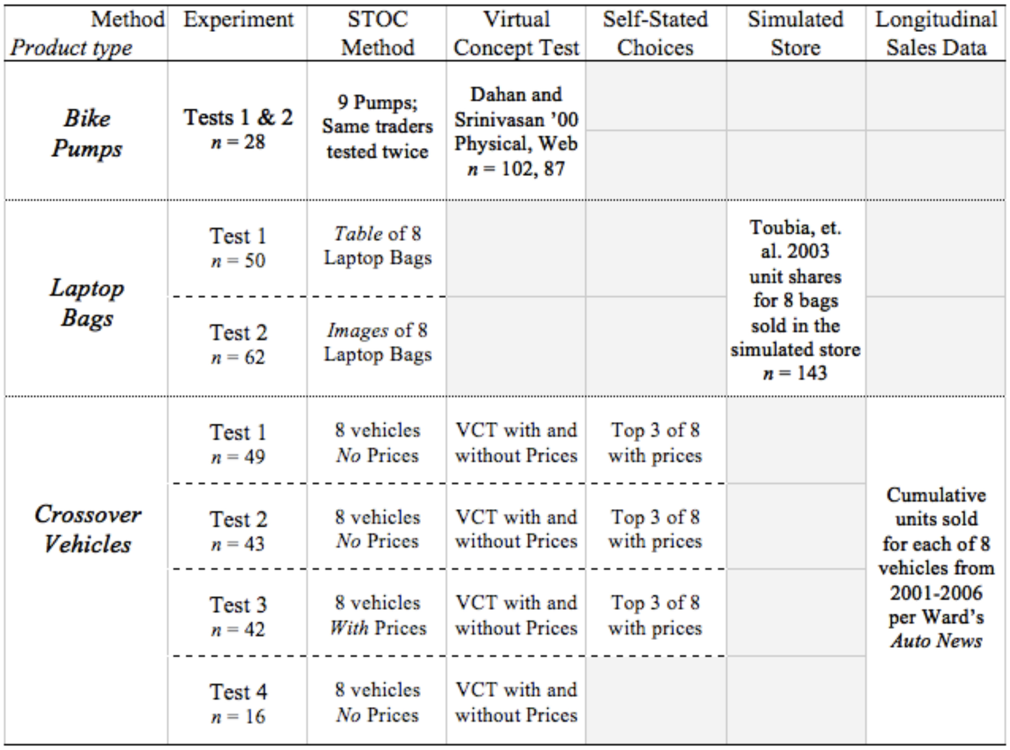
Concept testing enables new product development teams to identify those concepts most preferred by consumers. Dahan and Srinivasan (2000) present a virtual concept testing (VCT) methodology employing the Internet with virtual prototypes in place of real, physical ones. The goal of their study is to identify the most preferred of nine bicycle bump concepts versus the two commercially available products depicted in Figure C-2. The authors find that static images of the bike pumps on the Internet produce market share predictions that closely resemble those for real physical prototypes examined in person. We employ their physical prototype and static web virtual concept test results for bicycle pumps in hopes of validating the STOC method. Both bicycle pump STOC tests, conducted on the other side of the country and six years after Dahan and Srinivasan collected their data, were conducted with the same group of traders as a method of confirming test-to-test reproducibility as well as external validity.

Dahan and Hauser (2002) add multiple web-based market research methods to the mix, applying them to the eight existing and yet-to-be-released crossover vehicles depicted in Figure C-3. They also demonstrate a high degree of correlation between web-based concept testing and respondents self-stated-choices as measured by simple surveys. We test STOC in four independent crossover vehicle experiments and compare our results against self-stated data in three of them and against virtual concept testing (VCT) in all four. We estimate VCT preferences at the individual level in two ways: including both product preferences and vehicle prices to determine each traders utility score, and utilizing product preferences alone, excluding the effect of vehicle prices. In each case, we then aggregate individual preferences to generate market share estimates.

Toubia et al. (2003) develop a new polyhedral adaptive approach to conjoint analysis, and test it against existing adaptive and static conjoint methods using the example of customizable laptop PC messenger bags sold for real money through a simulated store. Their work demonstrates the effectiveness of their method, but more importantly for the present research offers an excellent data set for validating STOC. We focus on eight randomly chosen bags, representing a range of popularity (market share) actually sold to 43% of the respondents in their research. Two STOC tests were run to measure preferences for the same eight bags, but utilizing two different forms of stimuli: the table shown in Figure C-4 and the individual images shown in Figure C-6.

Additionally, in the six years following the crossover vehicle STOC tests, that is from 2001-2006, we also collected unit sales data for each of the eight vehicles from Wards Automotive News. These data are used as a test of external validity and the predictive power of the STOC method.

We are grateful for the cooperation of the aforementioned researchers who enabled us to adopt the identical product concept illustrations in our STOC tests. Thus, we are able to compare results for identical market research problems using STOC versus each of the prior methods. We attempt to validate our method in the eight STOC trading tests conducted from 2000 to 2002, as summarized in Table C.1. Traders in the first six tests were MBA students, but additional tests included attendees from the MIT Center for Electronic Business conference (crossover vehicle test 3) and more senior managers attending executive education classes (crossover vehicle test 4). All eight tests were run under controlled conditions in a business school trading laboratory.



Method *Product type*	Experiment	STOC Method	Virtual Concept Test	Self-Stated Choices	Simulated Store	Longitudinal Sales Data
Bike Pumps	**Tests 1 & 2** ***n* = 28**	**9 Pumps; Same traders tested twice**	**Dahan and Srinivasan '00 Physical, Web** ***n* = 102, 87**			
Laptop Bags	Test 1 *n* = 50	*Table* of 8 Laptop Bags			**Toubia, et. al. 2003 unit shares for 8 bags sold in the simulated store** ***n* = 143**	
	Test 2 *n* = 62	*Images* of 8 Laptop Bags				
Crossover Vehicles	Test 1 *n* = 49	8 vehicles *No* Prices	VCT with and without Prices	Top 3 of 8 with prices		**Cumulative units sold for each of 8 vehicles from 2001-2006 per Ward's *Auto News***
	Test 2 *n* = 43	8 vehicles *No* Prices	VCT with and without Prices	Top 3 of 8 with prices		
	Test 3 *n* = 42	8 vehicles *With* Prices	VCT with and without Prices	Top 3 of 8 with prices		
	Test 4 *n* = 16	8 vehicles *No* Prices	VCT with and without Prices			

Table C.1: Data Collected for each Product Category

For each of the above tests, STOC user interface was employed as shown in Figure C-7, and each test ran in under one hour including instructions and wrap up.

C.2.2 Prediction Markets

Non-financial *prediction markets* have been established for political elections, movie box office estimation, and other real world outcomes. The Iowa Electronic Markets (IEM) pioneered prediction markets for the purpose of forecasting election results (Forsythe et al., 1993). The IEM was founded for research and educational purposes. Trading profits from the market provide incentives for traders to collect and process information about future events. The IEM features real-money futures markets in which contract payoffs depend on the outcome of political and economic events. IEMs predictions have outperformed most national polls. Similarly, the Hollywood Stock Exchange, HSX.com, has provided accurate predictions of movie box office results (Spann and Skiera, 2003). The Foresight Exchange (FX) , predict the probability of future events occurring such as changes in the environment, scientific breakthroughs, the collapse of companies, or political and news outcomes. Companies such as Hewlett Packard, Microsoft, Best Buy and Google have employed prediction markets to forecast printer sales, software release dates, consumer electronics sales, and software take-up rates.

Prediction markets share with STOC the benefits of information aggregation, the joy of competitive play, the ability to learn from others and the incentive to be accurate. Prediction markets focus on actual outcomes, operate for weeks, months, and sometimes years, and incorporate private information and news as it happens. STOC markets, in contrast, focus on concepts that may never come into existence, and therefore may never have actual outcomes, run for 10-60 minutes typically, and are not influenced by outside news. In fact, the only information available to STOC traders is the personal preferences they hold, their expectations of others preferences, and whatever they learn by observing STOC price movements.

C.2.3 Rational Expectations (RE) Models and Experimental Markets

Our trading experiments are closely related to the literatures in rational expectations (RE) models with asymmetric information and experimental markets. In a standard asymmetric information RE model (Grossman, 1981), heterogeneous agents with diverse information trade with each other and, under certain conditions, the market will converge to an equilibrium in which prices fully reveal all relevant information. The most important criterion for convergence is that agents condition their beliefs on market information. In particular, agents make inferences from market prices and quantities about other agents private information.

The RE model has received considerable attention in the study of experimental markets (Plott and Sunder, 1982, 1988; Forsythe and Lundholm, 1990; Davis and Holt, 1993). Studies of the informational efficiency of a market relative to the RE benchmark fall into two categories: markets with fully informed agents (insiders) and uninformed agents, and markets with many partially informed agents. In various experimental markets with human subjects, the results for both market structures are the same: markets eventually converge to the RE equilibrium, i.e., information aggregation and dissemination occur successfully.

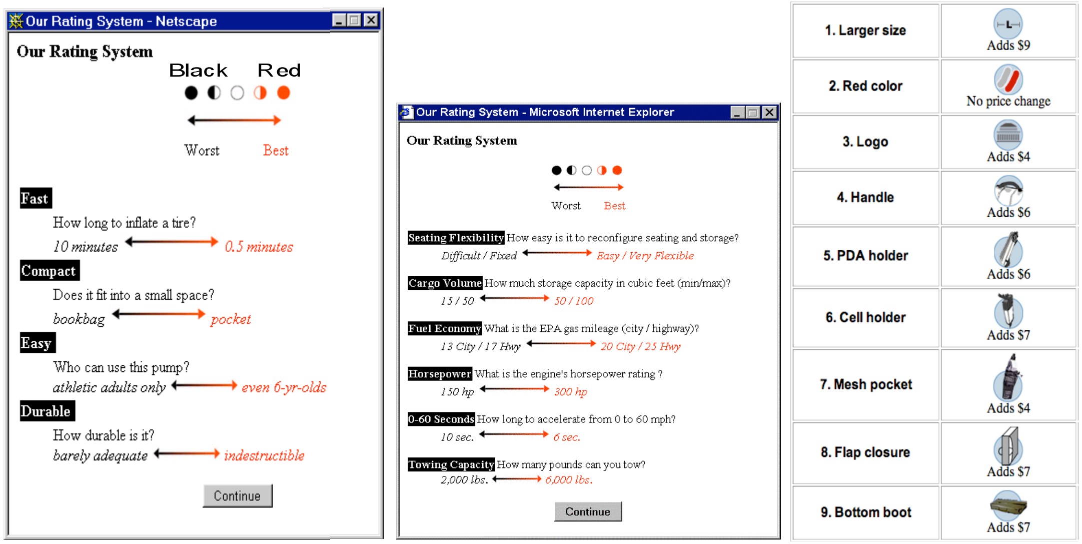
STOC trading share some characteristics with such experimental economics markets, and information aggregation and dissemination provide compelling explanation for the success of our STOC market. For example, traders who possess superior information about the products or have high confidence in their beliefs can be considered insiders. On the other hand, traders who have little knowledge or opinion of the products can be regarded as the uninformed. The interaction between the insider and uninformed constitutes information dissemination. What is intriguing about this scenario is that even when a subset of traders ignores the underlying product information and only focuses on market information, the market still converges to efficient prices that aggregate all the relevant information and beliefs.

Alternatively, individual traders may form their own beliefs about the products, acknowledging that market prices will depend on aggregate beliefs. This is similar to the information aggregation scenario in which there are no insiders, but where all traders are partially informed. Even in this case, where no single trader has full information, an RE equilibrium will be reached under very general conditions (Grossman, 1981; Davis and Holt, 1993, Chapter 7).

However, there is one important difference between our STOC market and the other exchanges in the experimental markets literature. In a typical experimental market, subjects preferences and their information set are fixed and assigned by the researchers. Therefore, even before trading begins, theoretical equilibrium prices can be calculated. In contrast, in a STOC market, neither the subjects preferences nor their information sets are knownin fact, these are what STOC market trading experiments are meant to discover. This suggests an important practical consideration in implementing STOC markets: the composition of traders should match the population of target consumers as closely as possible, or at least include traders with insight into the preferences of these consumers. For example, if the target population for a particular product is teenage female consumers, a STOC market consisting of middle-age males may not yield particularly useful preference rankings for that product. However, if the cross section of traders in a STOC market is representative of the target population, the force of market rationality will ensure that the price-discovery mechanism will provide an accurate measure of aggregate preferences.

C.3 Design of Markets and Securities

In addition to presenting full product concepts, we educated traders about the product attributes and attribute levels using Consumer Reports-style ratings as in Figure C-1.



Our Rating System - Netscape

Our Rating System

Black Red

● ◐ ○ ◑ ●

Worst Best

Fast

How long to inflate a tire?

10 minutes ⟷ *0.5 minutes*

Compact

Does it fit into a small space?

bookbag ⟷ *pocket*

Easy

Who can use this pump?

athletic adults only ⟷ *even 6-yr-olds*

Durable

How durable is it?

barely adequate ⟷ *indestructible*

Continue

Our Rating System - Microsoft Internet Explorer

Our Rating System

● ◐ ○ ◑ ●

Worst Best

Seating Flexibility How easy is it to reconfigure seating and storage?

Difficult / Fixed ⟷ *Easy / Very Flexible*

Cargo Volume How much storage capacity in cubic feet (min/max)?

15 / 50 ⟷ *50 / 100*

Fuel Economy What is the EPA gas mileage (city / highway)?

13 City / 17 Hwy ⟷ *20 City / 25 Hwy*

Horsepower What is the engine's horsepower rating ?

150 hp ⟷ *300 hp*

0-60 Seconds How long to accelerate from 0 to 60 mph?

10 sec. ⟷ *6 sec.*

Towing Capacity How many pounds can you tow?

2,000 lbs. ⟷ *6,000 lbs.*

Continue

1. Larger size	Adds $9
2. Red color	No price change
3. Logo	Adds $4
4. Handle	Adds $6
5. PDA holder	Adds $6
6. Cell holder	Adds $7
7. Mesh pocket	Adds $4
8. Flap closure	Adds $7
9. Bottom boot	Adds $7

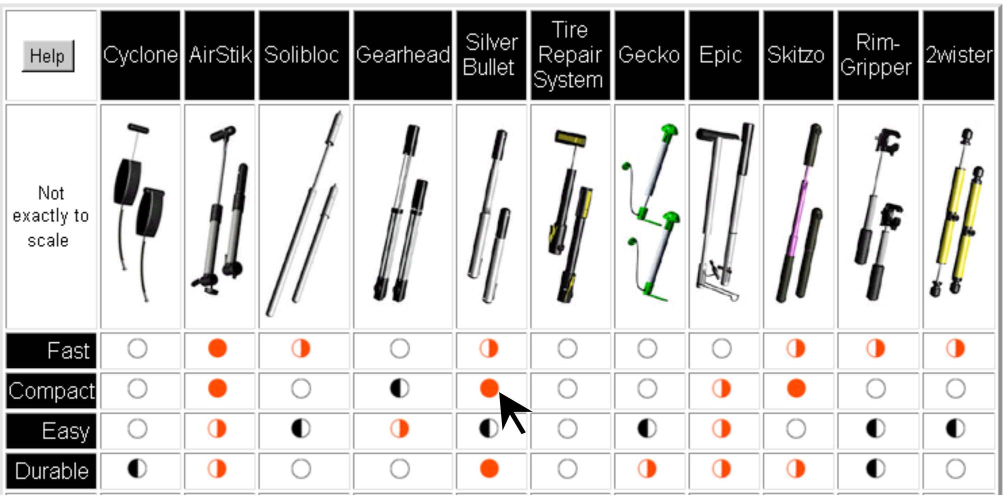
Figure C-1: Attributes for Bike Pumps, Crossover Vehicles and Laptop PC Messenger Bags

In our eight tests, the STOC method is applied to the product concepts described in Dahan and Srinivasan (2000) (Figure C-2), Dahan and Hauser (2002) (Figure C-3) and Toubia et al. (2003) (Figure C-4).

In order to anchor the value of the fictitious currency in the case of bike pumps, one of the eleven securities – Cyclone – has its price fixed at $10 and is not traded. Thus, Cyclone serves as a reference price or numeraire security. For example, if a trader thinks that TRS is worth twice as much as Cyclone, he or she would pay up to $20 for one share of TRS. The stocks of the ten freely traded concepts may be priced at any level, depending on the supply and demand in the market, i.e. the willingness of at least one trader to buy at the price at which another trader is willing to sell.

The eight crossover vehicles in Figure 3 were tested in 2000 and 2001 and consisted of three already released at the time (Lexus, Mercedes and BMW) and five yet-to-be-released vehicles (Pontiac, Acura, Buick, Audi and Toyota).

The eight laptop PC messenger bags shown in Figure 4 and Figure 5 were part of the controlled study described in Toubia et al. (2003) in which 330 first year MBA students were provided cash towards the purchase of a customized laptop PC messenger bag. These

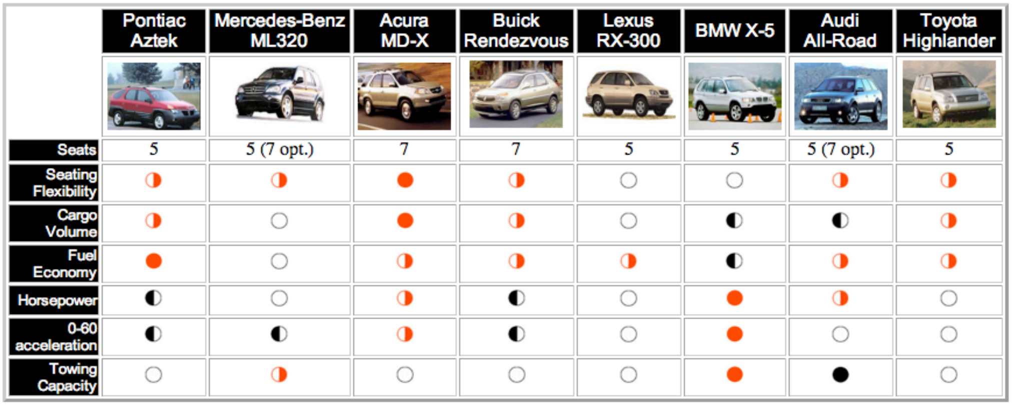


Help	Cyclone	AirStik	Solibloc	Gearhead	Silver Bullet	Tire Repair System	Gecko	Epic	Skitzo	Rim-Gripper	2wister
Not exactly to scale											
Fast	○	🟠	◑	○	◑	○	○	○	◑	◑	◑
Compact	○	🟠	○	◐	🟠	○	○	◑	🟠	○	○
Easy	○	◑	◐	◑	◐	○	◐	◑	○	◐	◐
Durable	◐	◑	○	○	🟠	○	◑	◑	◑	◐	○

● ◐ ○ ◑ 🟠

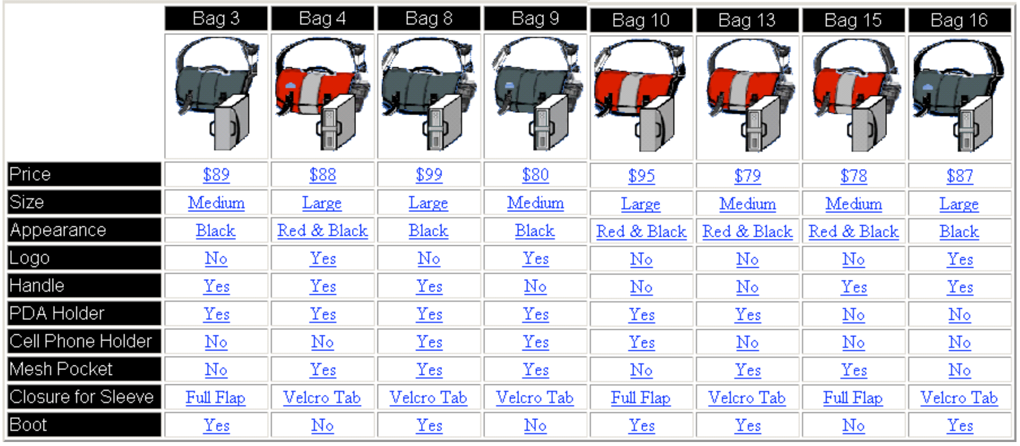
Worst ⟷ Best

Figure C-2: 11 Bike Pump Product Concepts Underlying the STOC Securities



	Pontiac Aztek	Mercedes-Benz ML320	Acura MD-X	Buick Rendezvous	Lexus RX-300	BMW X-5	Audi All-Road	Toyota Highlander
Seats	5	5 (7 opt.)	7	7	5	5	5 (7 opt.)	5
Seating Flexibility	◑	◑	🟠	◑	○	○	◑	◑
Cargo Volume	◑	○	🟠	◑	○	◐	◐	◑
Fuel Economy	🟠	○	◑	◑	◑	◐	◑	◑
Horsepower	◐	○	◑	◐	○	🟠	◑	○
0-60 acceleration	◐	◐	◑	◐	○	🟠	○	○
Towing Capacity	○	◑	○	○	○	🟠	●	○

Figure C-3: 8 Crossover Vehicles



	Bag 3	Bag 4	Bag 8	Bag 9	Bag 10	Bag 13	Bag 15	Bag 16
Price	$89	$88	$99	$80	$95	$79	$78	$87
Size	Medium	Large	Large	Medium	Large	Medium	Medium	Large
Appearance	Black	Red & Black	Black	Black	Red & Black	Red & Black	Red & Black	Black
Logo	No	Yes	No	Yes	No	No	No	Yes
Handle	Yes	Yes	Yes	No	No	No	Yes	Yes
PDA Holder	Yes	Yes	Yes	Yes	Yes	Yes	No	No
Cell Phone Holder	No	No	Yes	Yes	Yes	No	No	No
Mesh Pocket	No	Yes	Yes	Yes	No	Yes	Yes	No
Closure for Sleeve	Full Flap	Velcro Tab	Velcro Tab	Velcro Tab	Full Flap	Velcro Tab	Full Flap	Velcro Tab
Boot	Yes	No	Yes	No	Yes	Yes	No	Yes

Figure C-4: 8 Laptop PC Messenger Bags

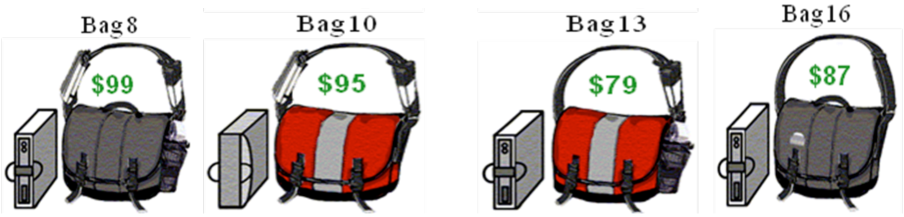


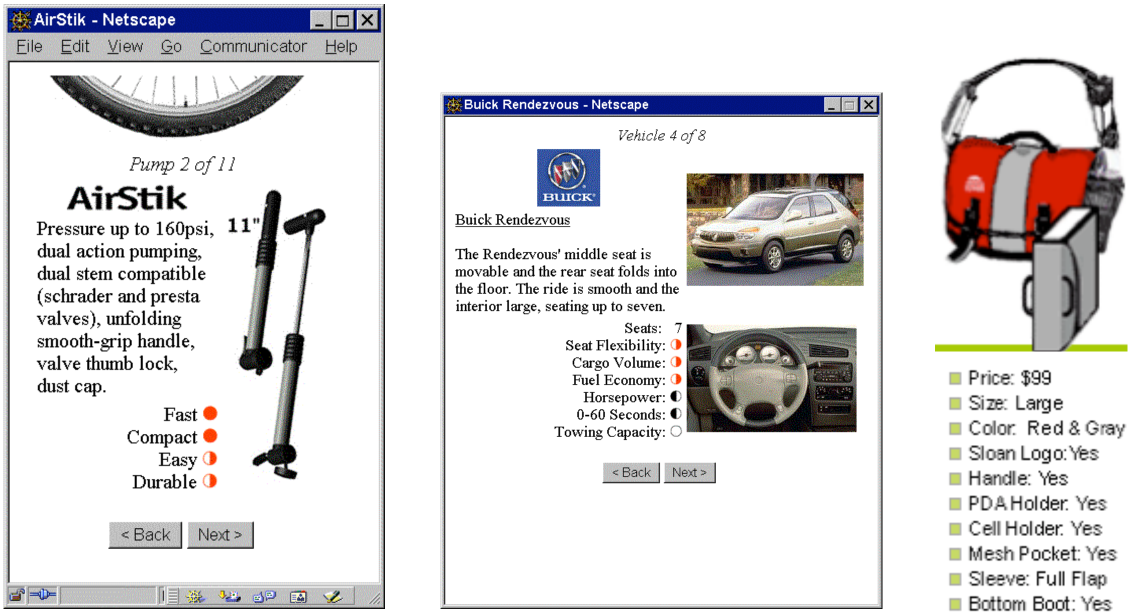
Figure C-5: Laptop PC Messenger Bags Depicted Without the Attribute Data Table

eight bags include designs that ranged from low to medium to high popularity amongst the 143 respondents to the original study who bought them from a simulated store with actual cash.

In the first laptop bag STOC test, traders saw the eight laptop bags in the table shown in Figure C-4. In test 2, the eight laptop bags were depicted as simpler images, four of which are reproduced in Figure C-6, leaving out the table of product attributes and simply showing nine product attributes visually rather than verbally. The eight laptop PC messenger bags depicted in the two types of experiments are identical.

The reason that a new set of stimuli were created to represent the same eight bags, frankly, was that the tabular form of presenting the bags was not well-received nor understood by the traders. In section C.5, this will become more apparent when the trading results of the two STOC tests are analyzed.

Each stock security represents a particular bike pump, crossover vehicle, or laptop PC messenger bag. The objective of traders in the STOC game is to maximize the value of ones


AirStik - Netscape
File Edit View Go Communicator Help
< Back
Next >

Pump 2 of 11

AirStik

Pressure up to 160psi, dual action pumping, dual stem compatible (schrader and presta valves), unfolding smooth-grip handle, valve thumb lock, dust cap.

11"

Fast ●
Compact ●
Easy ◑
Durable ◑

Buick Rendezvous - Netscape
BUICK
< Back
Next >

Vehicle 4 of 8

Buick Rendezvous

The Rendezvous' middle seat is movable and the rear seat folds into the floor. The ride is smooth and the interior large, seating up to seven.

Seats: 7
Seat Flexibility: ◑
Cargo Volume: ◑
Fuel Economy: ◑
Horsepower: ◐
0-60 Seconds: ◐
Towing Capacity: ○

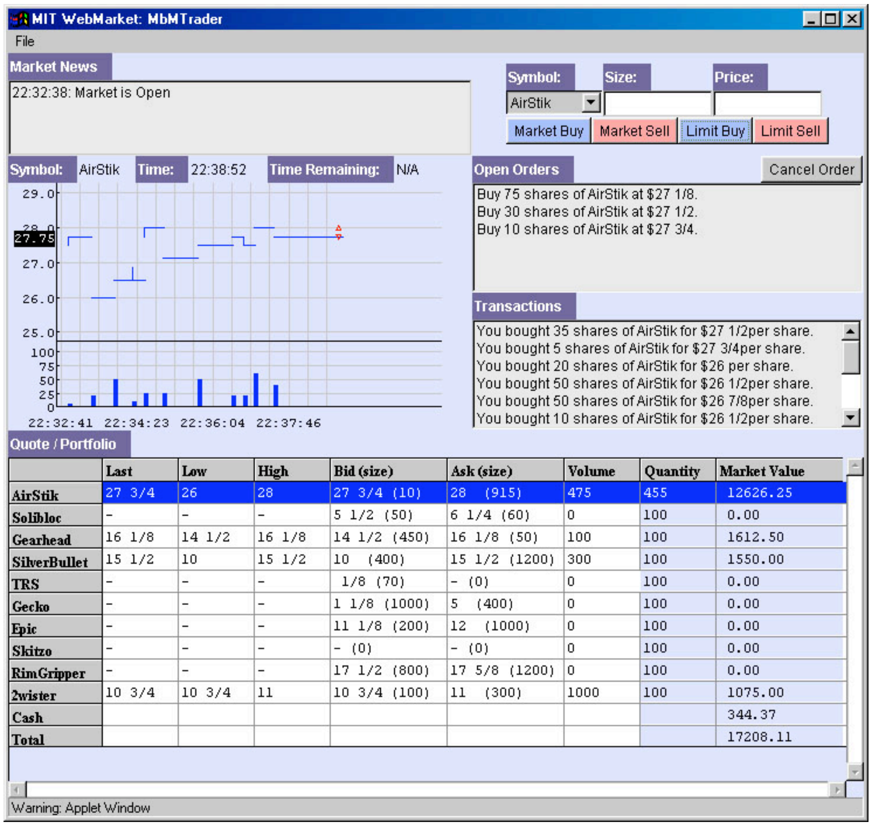
- Price: $99
- Size: Large
- Color: Red & Gray
- Sloan Logo:Yes
- Handle: Yes
- PDA Holder: Yes
- Cell Holder: Yes
- Mesh Pocket: Yes
- Sleeve: Full Flap
- Bottom Boot: Yes

Figure C-6: Typical Product Information for Bike Pumps, Crossover Vehicles, and Laptop PC Messenger Bags

portfolio at market close. The value of a portfolio is calculated as the sum of the cash on hand plus the total market value of the stocks, as determined by the closing market price. Participants strive to maximize their profits by trading the stocks by buying low and selling high. Beyond the stimuli they are shown, the only available information on which to base trading decisions is (1) their own preferences for the product concepts, (2) their perceptions of others preferences, and any (3) information they can glean from the trading dynamics of the game. Fictitious money is used in the markets, but top players may be rewarded with actual prizes and recognition by their trading peers. This provides the participants an incentive to perform well in the experiments. In these eight tests, peer recognition was the only incentive offered.

Each trader is provided with an identical portfolio that consists of $10,000 of cash and 100 shares of each security. No borrowing or short-selling is permitted in the market. Participants trade the securities through the graphical user interface shown in Figure C-7.

Market information available to the traders includes the last transaction price and size, current bid/ask prices and sizes, and a historical price and volume chart for each security. A trader can submit either a limit or market order to trade, or cancel an outstanding order that has not been executed. The markets are typical double auction markets with no market-makers. A transaction occurs when a market or limit order matches with a limit



22:32:38: Market is Open

Buy 75 shares of AirStik at $27 1/8.
Buy 30 shares of AirStik at $27 1/2.
Buy 10 shares of AirStik at $27 3/4.

You bought 35 shares of AirStik for $27 1/2per share.
You bought 5 shares of AirStik for $27 3/4per share.
You bought 20 shares of AirStik for $26 per share.
You bought 50 shares of AirStik for $26 1/2per share.
You bought 50 shares of AirStik for $26 7/8per share.
You bought 10 shares of AirStik for $26 1/2per share.

	Last	Low	High	Bid (size)	Ask (size)	Volume	Quantity	Market Value
AirStik	27 3/4	26	28	27 3/4 (10)	28 (915)	475	455	12626.25
Solibloc	-	-	-	5 1/2 (50)	6 1/4 (60)	0	100	0.00
Gearhead	16 1/8	14 1/2	16 1/8	14 1/2 (450)	16 1/8 (50)	100	100	1612.50
SilverBullet	15 1/2	10	15 1/2	10 (400)	15 1/2 (1200)	300	100	1550.00
TRS	-	-	-	1/8 (70)	- (0)	0	100	0.00
Gecko	-	-	-	1 1/8 (1000)	5 (400)	0	100	0.00
Epic	-	-	-	11 1/8 (200)	12 (1000)	0	100	0.00
Skitzo	-	-	-	- (0)	- (0)	0	100	0.00
RimGripper	-	-	-	17 1/2 (800)	17 5/8 (1200)	0	100	0.00
2wister	10 3/4	10 3/4	11	10 3/4 (100)	11 (300)	1000	100	1075.00
Cash								344.37
Total								17208.11

Figure C-7: STOC Trading User Interface

order on the opposite side of the market. All prices are specified in sixteenths of a dollar.

The trading prices (lines) and volumes (bars) for a sample security in the two bike pump tests are shown in Figure C-8. We note that price converges to approximately $25 in both tests.

For analysis, we focus on the trading data, which consists of the time series of trading prices and quantities $(p_{1,i}, q_{1,i}), (p_{2,i}, q_{2,i}), ..., (p_{T_i,i}, q_{T_i,i})$, where i is the index for the i^{th} product and is the total number of the cleared trades for the i^{th} product. Our hypothesis is that prices reveal market consensus of relative preferences for each product concept. In particular, we propose that a products market share can be predicted by its relative STOC price. In addition to the closing price, we consider other metrics that take into account all transactions during the session: the high, low, mean, median and volume weighted average prices. The high, low, mean and median prices are calculated from the time series of trade

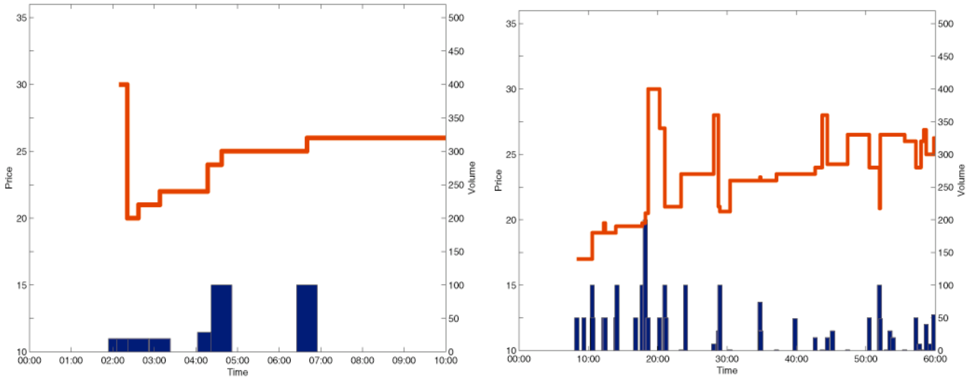


Figure C-8: Price and Volume History of AirStik for Test 1 (left) and Test 2 (right).

prices $p_{1,i}, p_{2,i}, ..., p_{T_i,i}$; the volume-weighted average price (VWAP) is computed as follows:

$$VWAP_i = \frac{\sum_{t=1}^{T_i} p_{t,i} q_{t,i}}{\sum_{t=1}^{T_i} q_{t,i}}. \tag{C.1}$$

The mean, high and low prices are sensitive to outliers – a small number of transactions that occur at extreme prices. All but VWAP ignore the volume in a transaction and treat all trades equally. Volume can be regarded as a measure of the amount of information in a transaction. A trade with higher volume may well be more informative than one with lower volume, since traders are risking more when they trade larger quantities of a stock. In our markets, volume is also related to how confident the traders are at the corresponding transaction price. VWAP effectively summarizes the prices by considering the amount of information and confidence behind the trades. In practice, VWAP has been a widely accepted benchmark price in financial markets. It is a commonly used metric for the evaluation of trade executions. So we might expect VWAP to more fully capture the consensus preferences of the traders.

Now given a price statistic $\overline{p}_i$, which can be the high, low, closing, mean, median or volume weighted average prices, we can arbitrarily compute predicted market share as the relative market capitalization,

$$MS_i = \frac{\overline{p}_i n}{\sum_{j=1}^{N} \overline{p}_j n} = \frac{\overline{p}_i}{\sum_{j=1}^{N} \overline{p}_j}, \tag{C.2}$$

where N is the number of securities comprising the market and n is the total number of shares for a security, which is equal for each security (each trader is given an identical endowment of each security prior to trading). Among the four price statistics, we expect the median price and VWAP to be particularly robust against potential price volatility.

C.4 Possible Trading Strategies

Our market tests are intended to aggregate diverse preferences or beliefs from all traders. We have evidence that the traders individual preferences prior to the STOC game were quite heterogeneous as reflected in stated-choice surveys and virtual concept tests run prior to the start of each game.

Ones beliefs and preferences, and the trading strategy based upon them, may derive from three related elements:

1. Product Information. This is what a participant knows about the underlying products. All participants are provided with the same facts and specifications of the products, but they may have obtained extra product information from their personal experience outside the experiments.

2. Personal Preferences. This is what surveys and polls typically collect. Although the aggregate preference of the whole market is the object of interest, ones personal view and biases may contribute to trading decisions.

3. Assessments of Others Preferences. A participant may form opinions and expectations of what others think so as to make profitable trading decisions. This adds a significant element of gaming and competition to the STOC method.

To get a sense of traders strategies before and after playing the STOC game, we surveyed the 77 traders in the two crossover vehicle STOC games, as summarized in Figure C-9.

We note that trader attitudes are quite heterogeneous for all three questions, narrowing slightly after trading, but not shifting in a statistically significant way. The picture that forms is that traders strategies encompass both self preferences and expectations of others preferences. Traders expect their target prices for buying and selling will vary considerably throughout the game, even though no new outside information is added after the start of trading. We can therefore infer that traders expect to learn from each other through the

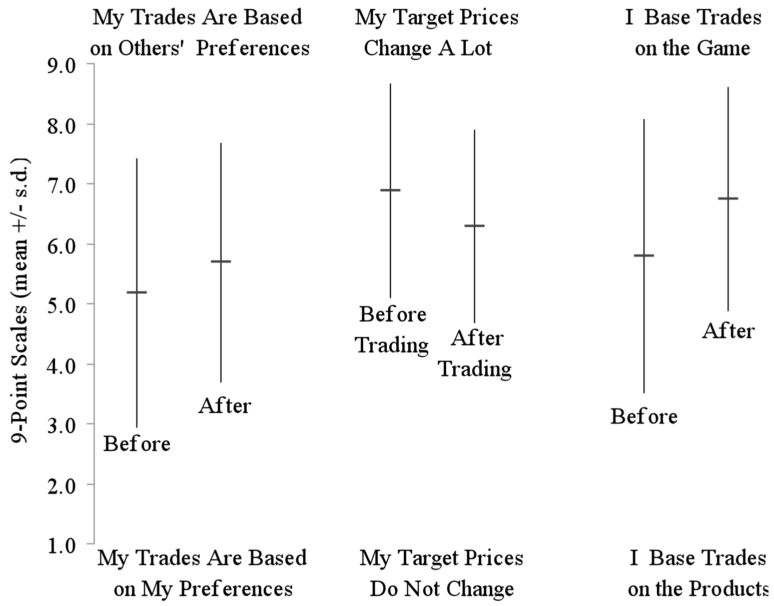


Figure C-9: Crossover Vehicle Trader Attitudes Before and After Trading ($n = 77$)

pricing mechanism. And traders focus slightly more on the gaming aspect of STOC than they do on the product concepts underlying the securities.

How are preferences aggregated in STOC markets? Not only do traders form their own assessment of value, but they also infer the stocks market value from the market itself. In typical experimental economics markets, both the preferences of the traders and the state of nature (for example, the probability distribution of a security payoff) are known to the researchers (Plott and Sunder, 1982, 1988; Forsythe and Lundholm, 1990; O'Brien and Srivastava, 1991). Traders are assigned preferences that specify securities payoffs in various possible states. The theoretical equilibrium (rational expectations equilibrium) prices can be derived given full information of the markets. The main focus of these experiments is whether and under what conditions rational expectations equilibria can be attained in double auction markets. Some attempts have been made to understand the convergence of prices and how learning occurs in the market as a whole. But it is unclear how individual human traders learn and react to the market. Attempts to model the trading strategies of individual traders from the market data may be overly ambitious. Below we try to shed some light on some possible strategies employed by different types of traders.

The objective of the trading game is to predict the final prices of the securities, trade accordingly, thereby maximizing ones final portfolio value. A trader may form an assessment of the fair values of the securities before trading begins. This assessment may naively take into account only her own preferences for the underlying products, or, if she is more sophisticated, what she perceives as the preferences of others. The trader then bases her trading decisions on her beliefs: she buys undervalued stocks and sells over-valued ones. During the course of trading, she may either maintain her valuations or update her beliefs in real time, conditioning on her observation of the market dynamics. Learning has taken place if the latter approach is adopted. But learning is a rather complex process because ones expectations of prices affect prices, prices are used to infer others assessments, and the inference of others assessments in turn affects both prices and expectations of prices.

Some traders may take a dramatically different approach by largely ignoring all fundamental information about the underlying products and focusing on stock market dynamics only. These traders play the roles of speculators or market-makers who try to gain from the market by taking advantage of price volatility, providing liquidity, or looking for arbitrage opportunities. Their presence may introduce mixed effects to the market. While they could enhance liquidity on one hand, they may also introduce speculative bubbles and excess volatility into the market.

In summary, STOC participants may include some combination of nave traders, long-term investors, and predatory arbitrageurs. The dynamics of the interactions between different groups within a given population is quite complex (Farmer and Lo, 1999; Farmer, 2002) and are beyond the scope of our study, but the principal of information revelation via the price-discovery process is the key to the STOC markets ability to infer aggregate preferences for concepts.

C.5 Results of STOC Tests

The outcomes of the eight STOC tests for three product categories which are summarized in Table C.1 led to our first key result regarding which metric best summarizes trading. In Figure C-10, we see that Volume-Weighted Average Prices (VWAP) fit the validation data better and more consistently than five alternative metrics.

In subsequent results, therefore, we utilize VWAP outcomes to check internal and external validity. Recall that VWAP summarizes all of the trades made from start to finish

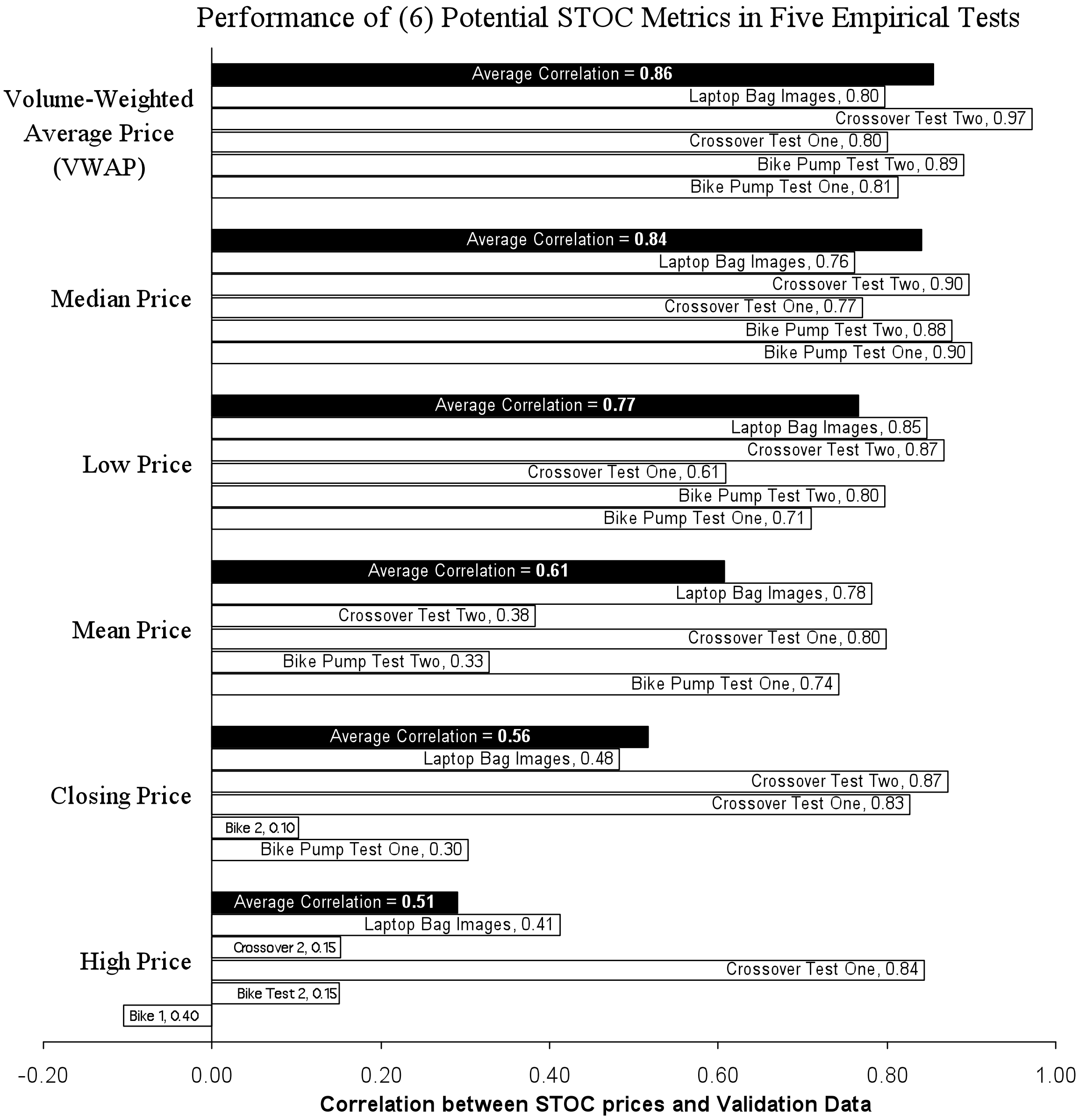


Figure C-10: Which STOC Metric Correlates Best to Validation Data?

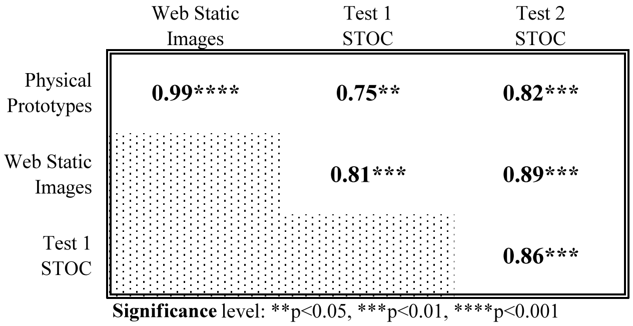
during a trading game, weighting each trade price by the number of shares traded. Closing and High prices, which do not depend on all trades but rather depend exclusively on the final or highest trade for each security, had the worst fits with the validation data. One potential cause of the poor correlation is that in early tests we allowed market orders, in which a trader does not specify a price, and is therefore vulnerable to executing a trade at an unreasonable price. After seeing this occur in approximately 1% of trades in our two bicycle pump tests, we eliminated the option of placing market orders and required each trader to specify a price by placing limit orders. We note that the VWAP and median metrics are less sensitive to a small number of trades at extreme prices.

Another explanation for the dramatic difference in correlation indicates that STOC markets, unlike prediction markets, behave more like traditional market research in which data is sampled from distributions with stationary means. In financial and prediction markets, this is not the case as security prices do not have stationary means due to the continuous arrival of new information. In our case, each STOC trade is similar to a survey response, and additional traders and greater trading time increase the sample size.

Shin and Dahan (2007) develop a statistical model to test whether STOC market data implies stationary on non-stationary security prices. They analyze the same trading data as in the present research, and employ unit-root tests to verify the stationarity or lack thereof of the mean prices for each security. The α coefficients in their model are highly correlated to our VWAP data, and their results support the conclusion that an ideal STOC metric should include all trades.

We wanted to see whether traders were consistent from test 1 to test 2, and whether the preference consensus represented by the STOC results matched those of Dahan and Srinivasan (2000). To verify the validity of the market method, we ask two questions: (1) whether the results from the market method are consistent across different experiments, and (2) how close the results from the markets are to those from Dahan and Srinivasan (2000).

We find that the top three products (Skitzo, Silver Bullet and Epic), in terms of predicted market share and rankings, are the same in the two experiments, as well as in the original Web Static data in the original Virtual Concept Test research. In a typical concept testing process, it is important to be able to identify the best designs so as to allocate resources to those opportunities with the greatest potential, and STOC seems to fulfill this role well.



	Web Static Images	Test 1 STOC	Test 2 STOC
Physical Prototypes	**0.99******	**0.75****	**0.82*****
Web Static Images		**0.81*****	**0.89*****
Test 1 STOC			**0.86*****

Significance level: **p<0.05, ***p<0.01, ****p<0.001

Table C.2: Correlations Between (2) Bicycle Pump STOC Tests and Validation Data From Dahan and Srinivasan (2000)

For consistency across experiments, we calculate a pair-wise sample Pearson correlation of 0.86 between market share predictions based on the VWAPs from the two tests. Comparing STOC with the original VCT web static data, correlations of 0.81 and 0.89, respectively, for tests 1 and 2 are slightly higher than those between STOC and the physical prototype results (0.75 and 0.82, respectively). This is consistent with the fact that the two groups of STOC traders were shown only the web static images and not the physical prototypes. Also noteworthy is that test-to-test reproducibility is quite high, with a correlation of 0.86.

The results from these initial bike pump experiments show a remarkable agreement with those from the Dahan and Srinivasan study despite fundamental differences between the two methods, timing and geography. Differences include those in the data collection mechanism (a virtual security market versus a virtual shopping experience), the modeling of the predicted market share (the use of relative security prices versus individual level conjoint analysis), the questions asked (what you prefer versus what the group prefers), and lastly the subject population (MIT students versus Stanford students).

The two STOC tests of laptop PC bags tell a slightly different, and quite remarkable story. The first laptop bag STOC test, the one in which the stimuli were presented in the form of a table with small images and feature details (Figure C-4) failed miserably in correlating to the simulated store sales. The second STOC test, however, the one in which traders learned about the eight laptop bags by viewing full-sized images without the feature table (Figure C-6) performed quite well, yielding a correlation of 0.80 with the simulated store data.

We attribute the dramatic difference in outcomes to the only factor that changed between

	Test 2 STOC Image Format	Simulated Store Sales
Test 1 STOC Table Format	-0.14	-0.05
Test 2 STOC Image Format		0.80**

Significance Level: **p < 0.05

Table C.3: Correlations Between (2) Laptop Bag STOC Tests and Simulated Store Unit Sales

the two tests, namely the stimuli. For the STOC method to perform well in capturing consensus preferences, traders must understand the product concepts reasonably well, so the quality of stimuli is crucial. While the feature table seems to have confounded or confused the traders, the full product images must have resonated with them. Extensive pre-testing of STOC stimuli is advised.

The most complete data set we analyzed comes from the crossover vehicle case and the four STOC tests conducted using those eight stimuli. Key correlation results are summarized in Table C.4, with significant results in bold, and all are calculated within each group of traders. *Self-stated* survey data represent the normed market shares for vehicles ranked by each individual in the top 3 out of eight choices. *VCT w/Prices* represent vehicle market shares based on scoring in the top three of eight vehicles using Dahan and Srinivasan (2000) methodology and accounting for the price of each vehicle when calculating utility. *VCT NO Prices* is the same calculation, but based only upon vehicle preferences without accounting for vehicle prices in the utility calculations. And, as before, "STOC" represents the normed marker shares based on the volume-weighted average prices of each of the eight securities.

Several significant results are captured in Table C.4, including the following five:

1. STOC vs. Actual Sales: The first row of the table, in which correlations to actual 2001-2006 unit sales of the eight vehicles were calculated for each method, reveals that all four STOC tests failed to predict actual sales. This result confirms our earlier analysis that STOC markets are not prediction markets, but rather measure a form of underlying preferences among the traders as we shall see shortly.

2. Self-Stated Choices and VCT w/Prices vs. Actual Share: Self-stated choices and

	Test 1 Self-Stated	Test 1 VCT w/Prices	Test 1 VCT NO Prices	Test 1 STOC	Test 2 Self-Stated	Test 2 VCT w/Prices	Test 2 VCT NO Prices	Test 2 STOC	Test 3 Self-Stated	Test 3 VCT w/Prices	Test 3 VCT NO Prices	Test 3 STOC	Test 4 VCT w/Prices	Test 4 VCT NO Prices	Test 4 STOC
Actual Units Sold 2001-2006	0.44	0.58	-0.2	0.22	0.42	**0.63***	-0.1	0.03	**0.62***	0.48	-0.0	0.52	0.52	-0.4	-0.4
Test 1 Self-Stated		0.54	0.54	**0.62***	**0.89*****	**0.79****	0.55	**0.64***	**0.91*****	0.37	**0.63***	**0.90*****	**0.74****	0.19	0.29
Test 1 VCT w/Prices			-0.1	-0.1	0.32	**0.91*****	-0.2	-0.1	0.55	**0.97******	-0.0	0.50	**0.94******	-0.3	-0.2
Test 1 VCT NO Prices				**0.80****	0.44	0.14	**0.96******	**0.91*****	0.51	-0.2	**0.95******	**0.62***	0.10	**0.81****	**0.72****
Test 1 STOC					0.58	0.20	**0.85*****	**0.92******	0.61	-0.2	**0.90*****	**0.69***	0.07	**0.65***	**0.62***
Test 2 Self-Stated						0.59	0.54	**0.66***	**0.84*****	0.12	0.48	**0.80****	0.49	0.00	0.07
Test 2 VCT w/Prices							0.08	0.20	**0.81****	**0.81****	0.27	**0.79****	**0.95******	-0.0	-0.0
Test 2 VCT NO Prices								**0.97******	0.52	-0.4	**0.93******	**0.62***	0.00	**0.76****	**0.66***
Test 2 STOC									0.65*	-0.3	**0.93******	**0.73****	0.07	**0.68***	0.57
Test 3 Self-Stated										0.35	0.61	**0.97******	**0.72****	0.12	0.07
Test 3 VCT w/Prices											-0.2	0.31	**0.85*****	-0.3	-0.2
Test 3 VCT NO Prices												**0.72****	0.19	**0.83*****	**0.73****
Test 3 STOC													**0.67***	0.31	0.19
Test 4 VCT w/Prices														-0.1	-0.0
Test 4 VCT NO Prices															**0.83****

Significance level: *p<0.10, **p<0.05, ***p<0.01, ****p<0.001

Table C.4: Correlations Between (4) Crossover Vehicle STOC Tests and Validation Data from Actual Unit Sales, Self-Stated Choices, and Virtual Concept Tests

Virtual Concept Testing with pricing did predict actual unit sales (correlations in the 0.42 to 0.63 range), though not in the statistical significance sense. One explanation for the superiority of these two measures over STOC is that there is an important difference between what people prefer, and what they are willing to pay for. STOC seems to zero in on preference rather than willingness-to-pay. Also, vehicle prices were not emphasized in STOC tests 1, 2 and 4, and were only featured prominently during STOC test 3, which was the only STOC test with some predictive value (0.52, but not significant).

3. STOC Test-to-Test Reproducibility: We saw with bike pumps that test-to-test reliability between STOC games using the same stimuli and same traders was quite good. In the crossover case, we can go further and measure test-to-test reliability across different groups of traders. Five of the six pairings of STOC tests reveal reasonably strong correlations between 0.57 and 0.92. But STOC tests 3 and 4 were not in agreement at all (correlation of 0.19), possibly because in STOC Test 3, where vehicle prices were emphasized, the higher-priced Audi, Mercedes and BMW vehicles garnered only 33% share. Test 4 had only 16 traders and vehicle prices were not emphasized, and

the three highest-priced vehicles garnered a whopping 64% share.

4. Stated Choice and VCT Test-Test Reproducibility: We note that the four VCT w/Price tests correlated amazingly well with each other (0.81 to 0.97) as did the VCT NO Price Tests (0.76 to 0.96). Similarly, aggregate Self-Stated data were highly correlated (0.84 to 0.91). So even though individual preferences were extremely heterogeneous, and group sample sizes were small (n = 16 to 49), aggregate preferences across groups were quite similar.

5. STOC vs. VCT with and Without Prices: STOC correlates remarkably well with the virtual concept test results when vehicle prices are not factored in (correlations of 0.80, 0.97, 0.72 and 0.83, respectively, for STOC tests 1 through 4). There was no correlation between the STOC tests and VCT with vehicle prices (-0.10 to 0.31). In short, STOC traders seem to focus on the vehicles when trading, but neither on the prices of those vehicles nor on the willingness-to-pay those vehicle prices.

We consider the degree of correlation within and across multiple tests and measures remarkable considering that most of the vehicles studied had not even entered the market and that the individuals comprising the trading groups were heterogeneous in their preferences and backgrounds.

C.6 Conclusions

In this paper we study a novel application of the market mechanism: the use of securities markets to aggregate and infer diverse consumer preferences. We implement this idea in the specific context of three product-concept testing studies that aim to predict potential market share for between eight to eleven product prototypes. The results from three seven of eight tests show remarkably high consistency among themselves, and significant correlation with independent preference measurement techniques. We note the importance of clear and salient stimuli, and the need for training and priming traders prior to the start of STOC games. We also caution that while the STOC methodology is particularly effective at screening most preferred concepts from among a larger set, it appears to be less effective at measuring price sensitivity or predicting actual sales. Of course, prediction markets can be designed to perform the latter, and other choice-based market research techniques are ideal for measuring price sensitivity.

The efficacy of STOC markets at identifying winning concepts may not be particularly surprising to economists. After all, Keynes (1958) commented on the similarities between stock selection and a beauty contest over a half-century ago:

> *... professional investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole ...*

The analogy is perhaps more accurate for describing what happens in financial markets and STOC games in the short run. After all, over the long run financial stock prices depend not only on investors subjective beliefs and expectations of others, but also on other objective information such as companies earning potentials and valuations of assets. On the other hand, the trading experiments presented in this paper are precisely beauty contests, since values of the virtual securities are derived endogenously from the preferences of the market participants, and their expectations of others preferences, both of which are largely subjective. To improve the reliability of STOC markets, one may need to anchor the values of the securities to some objective fundamental variables of the corresponding products. To test predictions of market shares, for example, one could compare security values with the realized market shares of the subset of existing products, or, barring the existence of real market share data, with the outcomes of actual customer choice surveys. We hope to refine STOC market methods along these lines in future research.

Bibliography

J. Angel. Limit versus market orders. Working Paper No. FINC-1377-01-293, School of Business Administration, Georgetown University, Washington, DC., 1994.

R. Battalio, J. Greene, B. Hatch, and R. Jennings. Does the order routing decision matter? Unpublished working paper, School of Business, Indiana University, Bloomington, IN., 1999.

L. E. Baum and J. A. Egon. An inequality with applications to statistical estimation for probabilistic functions of a markov process and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

L. E. Baum and G. R. Sell. Growth functions for transformations on manifolds. *Pacific Journal of Mathematics*, 27(2):211–227, 1968.

Y. Bengio and P. Frasconi. An input/output hmm architecture. *Advances in Neural Information Processing Systems*, 7:427–434, 1995.

Y. Bengio and P. Frasconi. Input-output hmm's for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249, 1996.

Yoshua Bengio. Markovian models for sequential data. *Neural Computing Surveys*, 2: 129–162, 1999.

J. Beran. *Statistics for Long-Memory Processes.* Chapman & Hall, 1994.

B. Biais, P. Hillion, and C. Spatt. An empirical analysis of the limit order book and the order flow in the paris bourse. *Journal of Finance*, 50:1655–1689, 1995.

J. P. Bouchaud, M. Mezard, and M. Potters. Statistical properties of stock order books: Empirical results and models. *Quantitative Finance*, 2:251–256, 2002.

J. P. Bouchaud, Y. Gefen, M. Potters, and M. Wyart. Fluctuations and reponse in financial markets: The subtle nature of 'random' price changes. *Quantitative Finance*, 4(2):176–190, 2004.

G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, 1970.

G. W. Burchill and C. H. Brodie. Voices into choices. Cambridge, MA: Center for Quality Management, 1997.

B. J. Calder. Focus groups and the nature of qualitative marketing research. *Journal of Marketing Research*, 14:353–364, 1977.

D. Challet and R. Stinchcombe. Analyzing and modeling $1 + 1d$ markets. *Physica A*, 300: 285–299, 2001.

D. Cox and D. Oakes. *Analysis of Survival Data.* Chapman & Hall, New York, 1984.

E. Dahan and J. R. Hauser. The virtual customer. *Journal of Product Innovation Management*, 19(5):332–353, 2002.

E. Dahan and V. Srinivasan. The predictive power of internet-based product concept testing using depiction and animation. *Journal of Product Innovation Management*, 17(2):99–109, 2000.

M. G. Daniels, J. D. Farmer, L. Gillemot, G. Iori, and E. Smith. Quantitative model of price diffusion and market friction based on trading as a mechanistic random process. *Physical Review Letters*, 90(10):108102, March 2003.

D. Davis and C. Holt. *Experimental Economics.* Princeton University Press, Princeton, NJ, 1993.

A.P. Dempster, N. M. Laird, and D. B. Rubin. Maximization-likelihood from incomplete data via em algorithm. *Journal of Royal Statistical Society B*, 39:1–38, 1977.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, January 1979.

B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589–599, December 1981.

Z. Eisler, J. Kertesz, F. Lillo, and R. N. Mantegna. Diffusive behavior and the modeling of characteristic times in limit order executions. *ArXiv Physics e-prints*, January 2007.

E. F. Fama. The behavior of stock market prices. *Journal of Business*, 38:34–105, 1965.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25:383–417, 1970.

J. D. Farmer. Market force, ecology and evolution. *Industrial and Corporate Change*, 11: 895–953, 2002.

J. D. Farmer and A. W. Lo. Frontiers of finance: Evolution and efficient markets. Number 96, pages 9991–9992, 1999.

J.D. Farmer and S. Mike. An empirical behavioral model of liquidity and volatility. Technical report, Santa Fe Institute Working Paper, 2006.

J.D. Farmer, L. Gillemot, F. Lillo, S. Mike, and A. Sen. What really causes large price changes? *Quantitative Finance*, 4(4):383–397, 2004.

E. F. Fern. The use of focus groups for idea generation: The effects of group size, acquaintanceship, and moderator on response quantity and quality. *Journal of Marketing Research*, 9:1–13, 1982.

R. Forsythe and R. Lundholm. Information aggregation in an experimental market. *Econometrica*, 58:309–347, 1990.

R. Forsythe, F. Nelson, G. Neumann, and J. Wright. The iowa presidential stock market: A field experiment. *Research in Experimental Economics*, pages 1–43, 1993.

M. Gell-Mann and C. Tsallis, editors. *Nonextensive Entropy - Interdisciplinary Applications.* SFI Studies in the Sciences of Complexity. Oxford University Press, 2004.

A. Gerig. *A Theory for Market Impact: How Order Flow Affects Stock Price.* PhD thesis, University of Illinois at Urbana-Champaign, 2007.

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

C. W. J. Granger and R. Joyeux. A introduction to long-memory time series and fractional differencing. *Journal of Time Series Analysis*, 1:15–29, 1980.

P. E. Green and V. Srinivasan. Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, pages 3–19, 1990.

P. E. Green and Y. Wind. New way to measure consumer judgment. *Harvard Business Review*, 53:107–117, July-August 1975.

A. J. Griffin and J. R. Hauser. The voice of the customer. *Marketing Science*, Winter:1–27, 1993.

S. J. Grossman. An introduction to the theory of rational expectations under asymmetric information. *Review of Economic Studies*, 48:541–559, 1981.

A.J. Hallinan. A review of the weibull distribution. *Journal of Quality Technology*, 25: 85–93, 1993.

P. Handa and R. Schwartz. Limit order execution. *Journal of Finance*, 51:1835–1861, 1996.

L. Harris and J. Hasbrouck. Market vs. limit orders: the superdot evidence on order submission strategy. *Journal of Financial and Quantitative Analysis*, 31:213–231, 1996.

F. Hayek. The use of knowledge in society. *American Economic Review*, XXXV(4):519–530, 1945.

B. Hollifield, R. Miller, and P. Sandas. An empirical analysis of pure limit order market. Working Paper, University of Pennsylvania, Philadelphia, PA, 1999.

J. R. M. Hosking. Fractional differencing. *Biometrika*, 68:165–176, 1981.

N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. A Wiley-Interscience Publication, New York, 2 edition, 1994.

J. Kalbfleisch and R. Prentice. *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, June 1958.

D. Keim and A. Madhavan. Anatomy of the trading process: Empirical evidence on the behavior of institutional traders. *Journal of Financial Economics*, 37:371–398, 1995.

J. M. Keynes. *The General Theory of Employment, Interest and Money.* Harcourt Brace, New York, 1958.

A. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, November 1985.

F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3):Article 1, 2004.

A.W. Lo, A.C. MacKinlay, and J. Zhang. Econometric models of limit-order executions. *Journal of Financial Economics*, 65:31–71, 2002.

LSEbulletin. Sets - four years on. Technical report, London Stock Exchange, 2001.

S. Maslov. Simple model of a limit order-driven market. *Physica A*, 278:571–578, 2000.

T. McInish and R. Wood. Hidden limit orders on the nyse. *Journal of Portfolio Management*, 21:19–26, 1986.

H. Mendelson. Market behavior in a clearing house. *Econometrica*, 50(6):1505–1524, 1982.

R. Miller. *Survival Analysis.* Wiley, New York, 1981.

J. O'Brien and S. Srivastava. Dynamic stock markets with multiple assets. *Journal of Finance*, 46:1811–1838, 1991.

M. Petersen and D. Fialkowski. Posted versus effective spreads. *Journal of Financial Economics*, 35:269–292, 1994.

S. Picoli, R.S. Mendes, and L.C. Malacarne. q-exponential, weibull, and q-weibull distributions: an empirical analysis. *Physica A*, 324:678–688, 2003.

C. R. Plott and S. Sunder. Efficiency of experimental security markets with insider information: An application of rational-expectations models. *Journal of Political Economy*, 90:663–698, 1982.

C. R. Plott and S. Sunder. Rational expectations and the aggregation of diverse information in laboratory security markets. *Econometrica*, 56:1085–1118, 1988.

Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–288, October 1967.

P. Samuelson. Proof that properly anticipated prices fluctuate randomly. *Industrial Management Review*, 6:41–49, 1965.

I. J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America*, 52(4):947–950, October 1964.

H. Shin and E. Dahan. A time-varying model of securities trading of concepts. UCLA Working Paper, 2007.

B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B*, 47(1):1–52, 1985.

E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy. Statistical theory of the continuous double auction. *Quantitative Finance*, 3(6):481–514, 2003.

V. L. Spann and B. Skiera. Internet-based virtual stock markets for business forecasting. *Management Science*, 49(10):1310–1326, 2003.

V. Srinivasan and A. D. Shocker. Estimating the weights for multiple attributes in a composite criterion using pairwise judgement. *Psychometrika*, 38(4):473–493, December 1973.

O. Toubia, D. I. Simester, J. R. Hauser, and E. Dahan. Fast polyhedral adaptive conjoint estimation. *Marketing Science*, 22(3):273–303, 2003.

C. Tsallis. Possible generalization of botzmann-gibbs statistics. *Journal of Statistical Physics*, 52:479, 1988.

G. L. Urban, J. R. Hauser, and J. H. Roberts. Prelaunch forecasting of new automobiles: Models and implementation. *Management Science*, 36(4):401–421, April 1990.

G. Wahba. Smoothing noisy data by spline functions. *Numerische Mathematik*, 24:383–393, 1975.