



THE UNIVERSITY OF  
**SYDNEY**

## **COPYRIGHT AND USE OF THIS THESIS**

This thesis must be used in accordance with the provisions of the Copyright Act 1968.

Reproduction of material protected by copyright may be an infringement of copyright and copyright owners may be entitled to take legal action against persons who infringe their copyright.

Section 51 (2) of the Copyright Act permits an authorized officer of a university library or archives to provide a copy (by communication or otherwise) of an unpublished thesis kept in the library or archives, to a person who satisfies the authorized officer that he or she requires the reproduction for the purposes of research or study.

The Copyright Act grants the creator of a work a number of moral rights, specifically the right of attribution, the right against false attribution and the right of integrity.

You may infringe the author's moral rights if you:

- fail to acknowledge the author of this thesis if you quote sections from the work
- attribute this thesis to another author
- subject this thesis to derogatory treatment which may prejudice the author's reputation

For further information contact the University's Director of Copyright Services

**[sydney.edu.au/copyright](http://sydney.edu.au/copyright)**

# Essays on Information Asymmetry and Price Impact in Market Microstructure



**Wei, Wang Chun**

B.Comm. (Honors 1st Class), Center for Actuarial Studies, The University of Melbourne, 2008  
Grad.Cert. Information Technology, School of Computer Science & Engineering, UNSW, 2013

presented to

the Discipline of Finance, Business School  
The University of Sydney, Australia

A thesis submitted in fulfillment of the requirements for the degree of  
*Doctor of Philosophy*

January, 2014

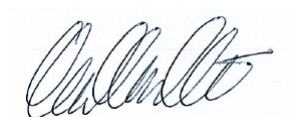
---

© Copyright by  
Wang Chun Wei  
2013

## Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other English or foreign examination board.

The PhD work was conducted from August 2010 under the supervision of Dr. Quan Gan at the University of Sydney.

A handwritten signature in black ink, appearing to read 'Wang Chun Wei', is written over a light blue rectangular background.

Wang Chun Wei,  
Sydney, Australia



## Abstract

This thesis comprises of topics on information asymmetry and price impact in market microstructure.

Our first paper introduces a new estimation method for the probability of informed trading - PIN (Easley, Kiefer, O'Hara and Paperman, 1996; Easley, Hvidkjaer and O'Hara, 2002; Easley, Hvidkjaer and O'Hara, 2010; Lin and Ke, 2011; and Yan and Zhang, 2012). PIN is an information asymmetry measure in market microstructure, and estimates the percentage of informed trading in the market. It is based on a structural model assuming Poisson arrival rates for informed and uninformed traders and daily Bernoulli probabilities on the occurrence of news, and type of news (e.g., good or bad news). We create a new method for estimating PIN using a hierarchical agglomerative clustering algorithm which we call Cluster PIN (CPIN). We show that it is superior to the most recent methods (Easley, Hvidkjaer and O'Hara, 2010; Lin and Ke, 2011; Yan and Zhang, 2012) in terms of accuracy, robustness and speed (approximately 300 times faster) and bypasses some of the problems faced with maximum likelihood estimation, such as the floating point exception. We show that CPIN is also able to explicitly classify trading days into 'good', 'bad' and 'no news' days which is not possible with existing approaches. This allows us to check the reliability of CPIN via an ex-post analysis of trading statistics (buy/sell volume, returns, volatility and spreads) under these three classification groups.

This thesis also examines price impact, which is used to measure the information content of trades. Hasbrouck (1991) states that trades convey information and the

magnitude of price impact for a given trade size is in proportion to the level of informed traders in the population. The price impact of a trade is estimated as cumulative quote revisions (or mid-price changes) due to incoming trades, i.e., signed log volume (Hasbrouck, 1991; Dufour and Engle, 2000). Hasbrouck (1991) use a bivariate VAR to model the interactions between quote revisions and trades, and show that lagged trades can impact quote revisions. Then the cumulative impulse response function (CIRF) of the VAR model is used to estimate the price impact of trades. Dufour and Engle (2000) show that both incoming trade duration and size can influence price impact as they reflect the level of informativeness of the trades.

Our second paper examines the drivers of quote revisions. We extend upon Dufour and Engle (2000) by also considering quoted spreads and depth as variables in the VAR model. From this, we show that quote revisions are not only affected by incoming trades, but also driven by order book illiquidity factors, such as quoted spreads and depth. Given the large number of parameters in our VAR model, we use adaptive lasso (Tibshirani, 1996; Zou, 2006; Hsu, Hung and Chang, 2008; and Ren and Zhang, 2010), to conduct robust variable selection and parameter estimation simultaneously; and show order book variables remain significant after variable selection.

We construct time-varying price impact by estimating our VAR model at weekly intervals from January 2007 to December 2012. To the best of our knowledge, our research is the first to analyze time-varying price impact. Our third paper examines the relationship between time-varying price impact and volatility. Our fourth paper studies the relationship between time-varying price impact and volume synchronized probability of informed trading - VPIN (see Easley, Lopez de Prado and O'Hara, 2012). Both measures relate to information asymmetry and risk aversion costs. However contrary to expectations, we find that there is a negative relationship between price impact and VPIN. We provide a heterogeneous rational expectation

equilibrium model to explain our empirical findings. We show the seemingly counter-intuitive result can be explained if one allowed for heterogeneity of beliefs amongst informed traders in processing news events.

## Acknowledgements

I would like to sincerely thank Dr. Quan Gan at the University of Sydney Business School for his teaching, support and advice on my PhD topic in financial modeling and econometric applications in market microstructure, and especially for introducing me to regularization, a topic more commonly found in statistics and machine learning. I would like to thank Prof. David Johnstone at the University of Sydney Business School for kindly advising on the third chapter of my PhD and Prof. Graham Partington who has taken time to address my numerous concerns. I acknowledge the financial support I received from the University of Sydney and the Australian Postgraduate Award. I appreciate the feedback I received from the participants at the Louis Bachelier Forum for Risk in Paris (2013), the American Committee for Asian Economic Studies Financial Econometrics Group in Melbourne (2012), the Australasian Finance and Banking Conference in Sydney (2012) and the Accounting and Finance Association Conference of Australia and New Zealand in Melbourne (2012).

Furthermore, I am indebted to my industry sponsor, Regal Funds Management, an equity hedge fund based in Sydney and Singapore, and in particular the quant team: Stephen Baldwin, Rajiv Thillainathan and George Mormanis, who helped me on my problems, technical or otherwise. I appreciate the numerous coffee runs with Jonathan Margo. Regal Funds Management has provided me with an excellent environment for learning and applying quantitative techniques in trading and has generously funded my graduate studies in computer science at UNSW which I studied concurrently alongside my PhD.

My gratitude also goes to the open source community, especially L<sup>A</sup>T<sub>E</sub>X and R, for generously making available programs and packages for research. I am also grateful for access to the Thomson Reuters Tick History database, without which research could not have been possible.

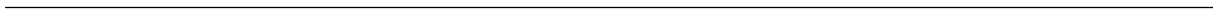
Most importantly, I like to sincerely thank my parents, Yuejin Wei and Yan Li, for their constant support in my attempts to gain greater knowledge in life.

*“The struggle itself towards the heights is enough to fill a man’s heart. One must  
imagine Sisyphus happy.”*

- Albert Camus<sup>1</sup>

---

<sup>1</sup>“I leave Sisyphus at the foot of the mountain. One always finds one’s burden again. But Sisyphus teaches the higher fidelity that negates the gods and raises rocks. He too concludes that all is well. This universe henceforth without a master seems to him neither sterile nor futile. Each atom of that stone, each mineral flake of that night-filled mountain, in itself, forms a world. The struggle itself toward the heights is enough to fill a man’s heart. One must imagine Sisyphus happy.” from *The Myth of Sisyphus and Other Essays* by Albert Camus



# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Market Microstructure . . . . .	1
1.2 Definitions . . . . .	3
1.2.1 The Definition of the <i>Trading Process</i> . . . . .	3
1.2.2 The Attributes of the Trading Process . . . . .	5
1.2.3 The Definition of <i>Information Asymmetry</i> . . . . .	5
1.2.4 The Definition of <i>Liquidity</i> . . . . .	6
1.2.5 The Definition of <i>Price Impact</i> . . . . .	7
1.3 The Topics Examined . . . . .	8
1.3.1 Topic 1: Using clusters to solve for PIN . . . . .	9
1.3.2 Topic 2: What drives quote revisions? The interactions between trades, quote revisions, durations, spreads and depth . . . . .	10
1.3.3 Topic 3: Price impact's relationship with volatility . . . . .	11
1.3.4 Topic 4: Price impact, VPIN and the role of informed trader heterogeneity	11
<b>References</b>	<b>13</b>
<b>2 Data</b>	<b>17</b>
2.1 Tick Data . . . . .	17
2.2 Data Preparation . . . . .	20
2.3 Trade Initiation . . . . .	21
2.3.1 Tick Rule . . . . .	22



## CONTENTS

---

2.3.2	Lee and Ready (1991) . . . . .	22
2.3.3	Ellis, Michaely and O'Hara (2000) . . . . .	22
2.4	The Korea Exchange . . . . .	22
2.4.1	Trading Rules . . . . .	23
2.4.2	Data Sample . . . . .	23
	<b>References</b>	<b>25</b>
<b>3</b>	<b>A Hierarchical Agglomerative Clustering approach for Estimating the Probability of Informed Trading</b>	<b>27</b>
3.1	Introduction . . . . .	28
3.2	Estimating PIN . . . . .	31
3.3	Cluster PIN . . . . .	33
3.4	A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations . . . . .	39
3.5	Employing CPIN as a starting value methodology for LK-PIN . . . . .	53
3.6	Classification of Good, Bad and No News Days . . . . .	55
3.7	Conclusions . . . . .	61
	<b>References</b>	<b>63</b>
<b>4</b>	<b>The Impact of Information Content and Illiquidity on Quote Revisions</b>	<b>65</b>
4.1	Introduction . . . . .	66
4.2	Literature Review . . . . .	71
4.3	Data . . . . .	74
4.4	The VARX model . . . . .	75
4.5	Regularization via Adaptive Lasso . . . . .	78
4.6	Adaptive Lasso Estimation . . . . .	80
4.7	Impulse Response Functions . . . . .	81
4.8	Empirical Implementation . . . . .	84
4.9	Empirical Results . . . . .	85
4.10	Conclusions . . . . .	90
4.11	Extensions on Price Impact . . . . .	90
	<b>References</b>	<b>97</b>

<b>5</b>	<b>The Price Impact of a Trade and its linkage with Volatility</b>	<b>103</b>
5.1	Introduction . . . . .	104
5.2	Price Impact and Kyle's $\lambda$ . . . . .	106
5.3	A Single Period Rational Expectations Equilibrium Model . . . . .	108
5.4	Empirical Tests . . . . .	111
5.5	Empirical Results . . . . .	114
5.6	Conclusions . . . . .	117
	<b>References</b>	<b>119</b>
<b>6</b>	<b>The Price Impact of a Trade, VPIN and the Role of Informed Trader Heterogeneity</b>	<b>123</b>
6.1	Introduction . . . . .	124
6.2	Data . . . . .	126
6.3	Volume Synchronized Probability of Informed Trading . . . . .	127
6.4	Price Impact of a Trade . . . . .	130
6.5	The Curious Case of Negative Correlation . . . . .	135
6.6	Empirical Results . . . . .	140
6.7	Theoretical Explanation . . . . .	143
6.7.1	Heterogeneous agents and continuous information flow . . . . .	144
6.7.2	Simulated Results . . . . .	151
6.8	Conclusion . . . . .	154
	<b>References</b>	<b>155</b>
<b>7</b>	<b>Conclusions</b>	<b>161</b>
	<b>References</b>	<b>163</b>
<b>A</b>	<b>Supplement for Chapter 4</b>	<b>165</b>

## CONTENTS

---

# List of Figures

1.1	The Dimenions of Liquidity (Muranaga, 1999) . . . . .	7
3.1	Net Order Flow Imbalance Histogram . . . . .	36
3.2	Net Order Flow Imbalance Dendrogram . . . . .	37
3.3	Buy and Sell Generation . . . . .	39
3.4	Test of Estimation Bias when increasing Trade Intensity . . . . .	43
3.5	Test of Estimation Variance when increasing Trade Intensity . . . . .	44
3.6	Test of Estimation Time when increasing Trade Intensity . . . . .	45
3.7	YZ-EHO-PIN Actual vs. Estimates . . . . .	46
3.8	YZ-LK-PIN Actual vs. Estimates . . . . .	47
3.9	CPIN Actual vs. Estimates . . . . .	48
3.10	YZ-EHO-PIN Estimates $k = 2,500$ vs $k = 5,000$ . . . . .	49
3.11	YZ-LK-PIN Estimates $k = 2,500$ vs $k = 5,000$ . . . . .	50
3.12	CPIN Estimates $k = 2,500$ vs $k = 5,000$ . . . . .	51
3.13	C-LK and YZ-LK Estimation Time Distribution . . . . .	54
4.1	Price Impact and Trade Impact: Cumulative Impulse Response Functions (CIRF) . . . . .	91
4.2	Price Impact and Trade Impact: Variability across Time . . . . .	92
4.3	Price Impact: Time-series Variability vs Mean . . . . .	94
4.4	Price Impact: Cross-sectional Variability vs. Mean . . . . .	95
5.1	Price Impact and Volatility: Samsung Electronics Scatterplot . . . . .	113
6.1	Time-series VPIN: Samsung Electronics . . . . .	130
6.2	Time-series Price Impact and Trade Impact . . . . .	134
6.3	Scatter-plot of Price Impact and VPIN . . . . .	135

## LIST OF FIGURES

---

6.4	An Illustration of Price Impact and VPIN from Samsung Electronics . . . . .	140
-----	---	-----

# List of Tables

1.1	Tick Trading Attributes . . . . .	5
2.1	Raw Transaction Data from Thomson Reuters Tick History . . . . .	19
2.2	Modified Transaction Data from Thomson Reuters Tick History . . . . .	20
2.3	Korea Daily Exchange Timetable . . . . .	24
2.4	Companies - Full List . . . . .	24
3.1	PIN Estimation Error for $k = 2,500$ and $k = 5,000$ . . . . .	52
3.2	Estimation Error between YZ-LK-PIN and C-LK-PIN . . . . .	54
3.3	Trading Behavior in Good New and Bad News Days (POSCO 005490 KS) . . . . .	58
3.4	Trading Behavior in Good New and Bad News Days (All stocks) . . . . .	59
3.5	Auto-regression on Classification (POSCO 005490 KS) . . . . .	60
3.6	Auto-regression on Classification (All Stocks) . . . . .	60
4.1	Summary on the VAR Quote Revision Equation . . . . .	86
4.2	Summary on the VAR Trade Equation . . . . .	87
4.3	Summary on the Factors Influencing Quote Revisions and Trades . . . . .	87
4.4	Average Coefficients for Samsung Electronics . . . . .	88
4.5	Average Coefficients for All Stocks . . . . .	89
4.6	Average Price Impact for All Stocks . . . . .	93
5.1	Linear Model: The Relationship between Volatility and Price Impact . . . . .	115
5.2	Linear Model: The Relationship between Volatility and Price Impact - continued	116
6.1	Pearson Correlations for Samsung Electronics . . . . .	137
6.2	Average Pearson Correlations for All Stocks . . . . .	139
6.3	Adaptive Lasso VAR Summary . . . . .	142

## LIST OF TABLES

---

6.4	Change of Information Flow Mean - $\mu_\eta$ . . . . .	151
6.5	Change of Information Flow Variance - $\sigma_\eta^2$ . . . . .	152
6.6	Change of Trader Confidence - $\sigma_c^2$ . . . . .	153
A.1	Adaptive Lasso Coefficients Samsung Electronics (005930 KS) . . . . .	166
A.2	Adaptive Lasso Coefficients Hyundai Motor (005380 KS) . . . . .	167
A.3	Average Adaptive Lasso Coefficients POSCO (005490 KS) . . . . .	168
A.4	Average Adaptive Lasso Coefficients Hyundai Mobis (012330 KS) . . . . .	169
A.5	Adaptive Lasso Coefficients Shinhan Financials Group (055550 KS) . . . . .	170
A.6	Adaptive Lasso Coefficients Kia Motors (000270 KS) . . . . .	171
A.7	Adaptive Lasso Coefficients SK Hynix (000660 KS) . . . . .	172
A.8	Adaptive Lasso Coefficients Hyundai Heavy Industries (009540 KS) . . . . .	173
A.9	Adaptive Lasso Coefficients KEPCO (015760 KS) . . . . .	174
A.10	Adaptive Lasso Coefficients SK Telecom (017670 KS) . . . . .	175
A.11	Adaptive Lasso Coefficients KT Corporation(030200 KS) . . . . .	176
A.12	Adaptive Lasso Coefficients KT&G Corporation (033780 KS) . . . . .	177
A.13	Adaptive Lasso Coefficients LG Chemicals (051910 KS) . . . . .	178
A.14	Adaptive Lasso Coefficients LG Electronics (066570 KS) . . . . .	179
A.15	Adaptive Lasso Coefficients Hana Financial Group (086790 KS) . . . . .	180
A.16	Adaptive Lasso Coefficients Hyundai Engineering & Construction (000720 KS) . . . . .	181
A.17	Adaptive Lasso Coefficients Samsung C&T Corporation (000830 KS) . . . . .	182
A.18	Adaptive Lasso Coefficients LG Corporation (003550 KS) . . . . .	183
A.19	Adaptive Lasso Coefficients Samsung Electro-Mechanics (009150 KS) . . . . .	184
A.20	Adaptive Lasso Coefficients Samsung Heavy Industries (010140 KS) . . . . .	185
A.21	Adaptive Lasso Coefficients S-Oil (010950 KS) . . . . .	186
A.22	Adaptive Lasso Coefficients LG Display (034220 KS) . . . . .	187
A.23	Adaptive Lasso Coefficients Kangwon Land Inc (035250 KS) . . . . .	188
A.24	Adaptive Lasso Coefficients LG Household & Health Care (051900 KS) . . . . .	189

# 1

## Introduction

### 1.1 Market Microstructure

Market microstructure has become an increasingly important and analyzed field within finance. It is fundamentally concerned with the details of how trading occurs within markets (see Madhavan, 2000; Harris, 2003). This dissertation comprises of topics on information asymmetry and price impact in market microstructure.

Whilst there exists a plethora of market microstructure theoretical models based upon information asymmetry, where market participants consist of informed and uninformed (liquidity) traders (Glosten and Milgrom, 1985; Kyle, 1985; Easley and O'Hara, 1987; Easley, Kiefer and OHara, 1997), less attention has been placed on developing models to estimate the level of information asymmetry in the market. One of the few measures for estimating the information asymmetry that have gained some traction in the literature is Easley, Kiefer, OHara and Paperman's (1996) probability of informed trading (PIN).

Our first paper (chapter 3) introduces a new estimation method for the probability of informed trading - PIN (Easley, Kiefer, O'Hara and Paperman, 1996; Easley, Hvidkjaer and O'Hara, 2002; Easley, Hvidkjaer and O'Hara, 2010; Lin and Ke, 2011; and Yan and Zhang, 2012). PIN is based on a structural model assuming Poisson arrival rates for informed and uninformed traders and daily Bernoulli probabilities on the occurrence of news, and type of news (e.g., good or bad news). We create a new method for estimating PIN using a hierarchical ag-



## 1. INTRODUCTION

---

glomerative clustering algorithm which we call Cluster PIN (CPIN). We show that it is superior to the most recent methods (Easley, Hvidkjaer and O'Hara, 2010; Lin and Ke, 2011; Yan and Zhang, 2012) in terms of accuracy, robustness and speed (approximately 300 times faster) and bypasses some of the problems faced with maximum likelihood estimation, such as the floating point exception. We show that CPIN is also able to explicitly classify trading days into 'good', 'bad' and 'no news' days which is not possible with existing approaches. This allows us to check the reliability of CPIN via an ex-post analysis of trading statistics (buy/sell volume, returns, volatility and spreads) under these three classification groups.

Analyzing price impact is central to understanding the price discovery process. We study price impact, which is used to measure the information content of trades. Both O'Hara (1995) and Dufour and Engle (2000) describe price discovery to be the mechanics of price formation - on how information is impounded into prices. This information impounding process is the crux of trading. Seminal works by Kyle (1985) and Glosten and Milgrom (1985) propose theoretical models showing how incoming order flow can cause price to change. Empirical works by Hasbrouck (1991), Dufour and Engle (2000), Pascual, Escibano and Tapia (2004), Escibano and Pasual (2006) show how trades can impact prices.

Hasbrouck (1991) states that trades convey information and the magnitude of price impact for a given trade size is in proportion to the level of informed traders in the population. The price impact of a trade is estimated as cumulative quote revisions (or mid-price changes) due to incoming trades, i.e., signed log volume (Hasbrouck, 1991; Dufour and Engle, 2000). Hasbrouck (1991) use a bivariate VAR to model the interactions between quote revisions and trades, and show that lagged trades can impact quote revisions. Then the cumulative impulse response function (CIRF) of the VAR model is used to estimate the price impact of trades. Dufour and Engle (2000) show that both incoming trade duration and size can influence price impact as they reflect the level of informativeness of the trades.

In our second paper (chapter 4), we examine the drivers of quote revisions. We extend upon Dufour and Engle (2000) by also considering quoted spreads and depth as variables in the VAR

model. From this, we show that quote revisions are not only affected by incoming trades, but also driven by order book illiquidity factors, such as quoted spreads and depth. Given the large number of parameters in our VAR model, we use adaptive lasso (Tibshirani, 1996; Zou, 2006; Hsu, Hung and Chang, 2008; and Ren and Zhang, 2010), to conduct robust variable selection and parameter estimation simultaneously; and show order book variables remain significant after variable selection.

Furthermore, we construct time-varying price impact by estimating our VAR model at weekly intervals from January 2007 to December 2012. To the best of our knowledge, our research is the first to analyze time-varying price impact. In our third paper (chapter 5), we show a positive relationship between time-varying price impact and volatility. In our fourth paper (chapter 6), we study the relationship between time-varying price impact and volume synchronized probability of informed trading - VPIN (see Easley, Lopez de Prado and O'Hara, 2012). Both measures relate to information asymmetry and risk aversion costs. However contrary to expectations, we find that there is a negative relationship between price impact and VPIN. We provide a heterogeneous rational expectation equilibrium models to explain our empirical findings. We show the seemingly counterintuitive result can be explained if one allowed for heterogeneity of beliefs amongst informed traders in processing news events.

Below we provide some definitions on the terminology and concepts used in this thesis. Then we provide a more detailed synopsis of the topics examined.

## 1.2 Definitions

### 1.2.1 The Definition of the *Trading Process*

O'Hara (1995) defines market microstructure to be the "*study of the process and outcomes of exchanging assets under a specific set of rules*", and trading to be the notion of exchanging assets. The Oxford English dictionary defines trading to be "*the action of buying and selling goods or the exchange (of something) for something else*". Whenever one discusses the concept

## 1. INTRODUCTION

---

of trading, invariably one talks about the exchange of assets; and implicit to the exchange, some form of valuation is necessary. The concept of valuation is fundamental to trading. We define trading on the stock exchange to be conceptually a mapping mechanism; and via trading, valuations are determined or learned. Put simply, through trading, anything can be priced. Information from various sources and formats (both tangible and intangible) are mapped onto a positive Real number line, i.e. the dollar traded price, by numerous transactions.

Trading: real world information  $\rightarrow \mathfrak{R}^+$

In valuation, an investment analyst values a stock by considering all the available information at his/her disposal about the stock and map it to the Real number line as a price.

Valuation: individual's information  $\rightarrow \mathfrak{R}^+$

Trading facilitates the discovery of the underlying price of the security because it involves the union of all analysts' information (and subsequent valuations), and their interactions (via trades).

Trading:  $\cup_{i \in \{1, \dots, N\}}$  individual's information<sub>*i*</sub>  $\rightarrow \mathfrak{R}^+$

where  $\cup_{i \in \{1, \dots, N\}}$  individual's information<sub>*i*</sub>  $\approx$  total real world information, which in part suggests Fama and French's efficient market hypothesis (EMH). We refer to semi-strong form efficiency, where public information available to all individuals would be instantly priced. However, insider information would be learned slowly by the market via trading between the interaction of more informed and less informed traders. This is loosely described as price discovery. To further extend this mapping concept, it is noted that prices are not the only attribute determined via trading, for example, volumes and depth. Hence more generally, for  $k$  attributes

Trading:  $\cup_{i \in \{1, \dots, N\}}$  individual's information<sub>*i*</sub>  $\rightarrow \mathfrak{R}^k$

We view trading to be a form of communication amongst rational profit maximizing individuals with heterogeneous information (i.e., beliefs). By communicating, market participants are able to learn new information and converge on the true underlying price of an asset. However,

this form of communication is not free. The cost attached for the uninformed trader is the chance of being taken advantage of by an informed trader. This interaction between informed and uninformed traders forms the basis of numerous theoretical and empirical papers in market microstructure, many of which we will explore in the chapters to come.

### 1.2.2 The Attributes of the Trading Process

Here we define some of the raw<sup>1</sup> attributes of the trading process mentioned above. These raw attributes are considered in our research based on transaction and order book data, but they are by no means exhaustive.

**Table 1.1:** Tick Trading Attributes

Attribute	Notation	Definition
Traded Price	$p_t$	The price per unit share agreed upon and executed for transaction $t$
Traded Volume	$v_t$	The number of shares agreed upon and executed for transaction $t$
Trade Durations	$d_t$	The time elapsed between transactions $t - 1$ and $t$
Best Bid Price	$q_t^{bid}$	The quote price of the highest bid in the order book before transaction $t$ and after $t - 1$
Best Bid Size	$v_t^{bid}$	The quote size of the highest bid in the order book before transaction $t$ and after $t - 1$
Best Ask Price	$q_t^{ask}$	The quote price of the lowest ask in the order book before transaction $t$ and after $t - 1$
Best Ask Size	$v_t^{ask}$	The quote size of the lowest ask in the order book before transaction $t$ and after $t - 1$

### 1.2.3 The Definition of *Information Asymmetry*

Information asymmetry is a product of heterogeneous beliefs and information sets. Unlike traditional works in finance, such as the Markowitz mean-variance CAPM framework, market microstructure practitioners do not believe in homogeneous information sets amongst market participants. Therefore, at any particular point in time, there exists both informed traders and uninformed traders (see Easley and O'Hara, 1992; Easley, Keifer, O'Hara and Paperman, 1996; Easley, Hvidkjaer and O'Hara, 2002). This discrepancy in information amongst traders is regarded as information asymmetry. Kyle (1985) describe the interactions between these two distinct groups are governed as follows: (a) informed traders want to maximize insider profits and (b) uninformed traders or market makers try to learn from the order flow and adjust quotes

<sup>1</sup>Raw meaning un-modified. For example, spreads are derived from bid and ask prices and therefore are derivative attributes of the trading process. Volatility is derived from transaction prices, and therefore is also a derivative attribute.

## 1. INTRODUCTION

---

respectively so that they are not taken advantage of. The larger the trades by the insiders, the more likely the uninformed traders will adjust the quotes more significantly; hence a quadratic programming problem arises for the insider to maximize profits. Several measures have been developed to measure the level of information asymmetry in the market. For example, one can measure the level of order flow imbalance between buy and sell initiated trades. In this dissertation we will examine two popular models that measure information asymmetry: the probability of informed trading i.e., PIN (see Easley, Keifer, O'Hara and Paperman, 1996; Easley, Hvidkjaer and O'Hara, 2002) and volume synchronized probability of informed trading i.e., VPIN (see Easley, Lopez de Prado and O'Hara, 2012).

### 1.2.4 The Definition of *Liquidity*

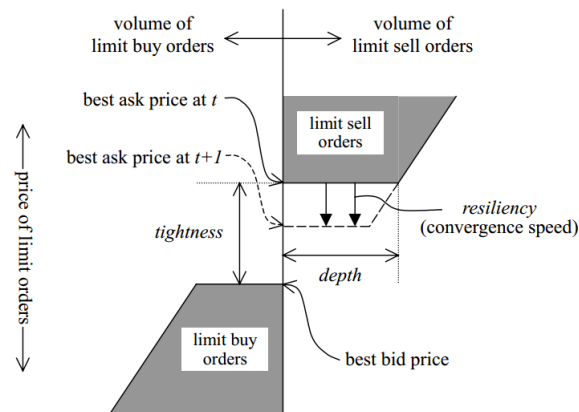
Liquidity is loosely defined as the ability to trade immediately at the price and volume you want. In markets where liquidity is low, you may not be able to transact at the price and volume you wish for, or may have to wait for a counter-party willing to trade. Invariably, transaction cost is higher when liquidity is low. Kyle (1985) defines market liquidity in terms of three concepts: *tightness*, *depth* and *resiliency*.

1. *Tightness* is defined to be the spread between the best bid and the ask price, i.e., the bid-ask spread
2. *Depth* is a market impact measure, it calculates the quote changes triggered by trade execution divided by the corresponding trade volume
3. *Resiliency* is the convergence speed of the bid-ask spread after trades

In Engle and Lange (1997) and Muranaga (1999), the practical aspects of market liquidity using Kyle's (1985) concepts are explained. *Tightness* explains the difference between the trade price and actual price. Active traders are required to cross the spread to execute their trades, tightness therefore is the cost of crossing that spread. *Depth* indicates the maximum number of stocks the order book can absorb at the current price level. It is measured by the average

between the size of the best bid quote and the size of the best ask quote. These two concepts defined by Kyle (1985) are easily measured via static metrics from the orderbook. *Resiliency*, the convergence speed after a large trade execution is harder to measure. Muranaga (1999) explains it as the elasticity of the orderbook. This can be measured as the slope of the order book - i.e. as we move up (move down) the ask side (bid side) of the order book, the cumulative increase in quote volume.

**Figure 1.1:** The Dimensions of Liquidity (Muranaga, 1999)



Muranaga and Shimizu (1999) argue that in order to examine liquidity, and its role in price discovery, not only static aspects such as bid-ask spreads or levels of order book depth should be considered. They suggest that liquidity can only be recognized during the dynamic process of trade execution. Dynamic indicators can show the actual result of trade execution. Bid-ask spreads may provide indication on liquidity conditions, however, it is by no means an estimate of market impact.

### 1.2.5 The Definition of *Price Impact*

Price impact is defined as the total and permanent change in price that is attributed to a trade shock., it can be described as a dynamic liquidity indicator. Price impact examines the magnitude quotes revise to incoming trades. To eliminate effects associated with the bid-ask bounce, it is normal convention to measure the midpoint price between the best bid and the

## 1. INTRODUCTION

---

best ask, rather than the transaction price. Therefore, it is not uncommon to consider the magnitude of quote revisions after a trade, rather than actual changes in transaction price. The methodology for estimating price impact is derived from Hasbrouck's (1991) seminal paper on the information content of stock trades. Quote revisions and trades are modeled in a bi-variate VAR model. The coefficients of the VAR model describe the lagged interactions between quote revisions and trades. For example, if we define quote revisions to be  $r_t$  and trades  $x_t$  (as per the notation of Hasbrouck, 1991), the standard bi-variate VAR is loosely specified as,

$$\begin{aligned} r_t &= \sum_{i=1}^p a_i r_{t-i} + \sum_{i=0}^p x_{t-i} + \nu_{1,t} \\ x_t &= \sum_{i=1}^p c_i r_{t-i} + \sum_{i=1}^p x_{t-i} + \nu_{2,t} \end{aligned} \tag{1.1}$$

We note the contemporaneous relationship between quote revisions and trades in the quote revision equation. This is because trades have an immediate impact on quote revision. The ordering is clear in our study, trades occur prior to quote revisions. Impulse responses are easy to interpret. Conceptually, we visualize running a unit trade shock  $\nu_{2,t}$  through the VAR system and work out its expected impact on  $r_{t+1}, r_{t+2}, \dots$ . As described in Hasbrouck (1991) and Dufour and Engle (2000), the estimation of  $\sum_{k=0}^{\infty} E(r_{t+k})$  provides an estimation of the price impact of a trade. Therefore, the cumulative impulse response of the bi-variate VAR model estimates the information content of trades. A higher quote revision to incoming trades is suggestive that the trades hold more information. In Escribano, Pascual and Tapia (2004), the cumulative impulse response functions are used as measures for risk aversion costs, as a higher quote revision is suggestive that the participants in the limit order book are more risk averse. Details on the price impact literature will be discussed in chapter 3.

### 1.3 The Topics Examined

Below we provide a synopsis of each topic in this dissertation. The underlying methodology used to estimate price impact in chapter 4 is employed across several subsequent chapters. References are provided at the end of each specific topic.

### 1.3.1 Topic 1: Using clusters to solve for PIN

Information asymmetry is an integral part of market microstructure, and forms the basis of many well-known theoretical models (see Glosten and Milgrom, 1985; Kyle, 1985; Easley and O’Hara, 1987; Easley, Kiefer and OHara, 1997). PIN (Easley, Kiefer, O’Hara and Paperman, 1996; Easley, Hvidkjaer and O’Hara, 2002) is a model developed to estimate the level of information asymmetry in the market. It is based on a structural model assuming Poisson arrival rates for informed and uninformed traders and daily Bernoulli probabilities on the occurrence of news, and type of news (e.g., good or bad news).

In topic 1, we study estimation techniques for PIN. We show that existing maximum likelihood estimation (MLE) techniques for estimating PIN are either inaccurate or time-consuming to compute. In earlier work by Easley, Kiefer, O’Hara and Paperman (1996) no discussion is provided on either employing an initial starting point algorithm, or on likelihood factorization. By ignoring these two points, estimation is inevitably inaccurate (Lin and Ke, 2011). We use Yan and Zhang’s (2012) initial value algorithm and compare the recent likelihood factorization approaches of Easley, Hvidkjaer and O’Hara (2010) and Lin and Ke (2011). We show that Lin and Ke (2011) is superior in terms of accuracy. In extension, we create our own methodology to estimate PIN that bypasses the computational issues associated with MLE. We show that our method involving hierarchical agglomerative clusters (HAC)<sup>1</sup> is more robust and 300x faster to estimate than both Lin and Ke (2011) and Easley, Hvidkjaer and O’Hara (2010). We also show that using HAC allows explicit classification of news days (into good news, bad news and no news days); this is not possible with MLE methods. We perform ex-post analysis on the explicit classifications to test the accuracy of our new method.

This topic is addressed in chapter 3.

---

<sup>1</sup>HAC for the remainder of this thesis refers to hierarchical agglomerative clustering. Readers are reminded not to be confused with the HAC in the Newey-West heteroskedasticity and autocorrelation-consistent estimator.



## 1. INTRODUCTION

---

### 1.3.2 Topic 2: What drives quote revisions? The interactions between trades, quote revisions, durations, spreads and depth

A fundamental question in market microstructure is: how is new information incorporated into prices? Hasbrouck's (1991) seminal paper shows that the level of price changes (permanent price shocks as measured via cumulative impulse response functions) depends on the sign and size of trades. Dufour and Engle (2000) use theory from Diamond and Verrecchia (1987) and Easley and O'Hara (1992) to examine the role durations (or time) have on price impact. They show that faster trading (shorter durations) meant more informed trading, i.e., higher price impact. We revisit Dufour and Engle's (2000) study with several extensions. Firstly, if trade durations have an impact on cumulative quote revisions, then is it possible for other trading attributes such as spreads and depths to also have an impact on quote revisions? Therefore, we extend upon Dufour and Engle (2000) by considering not only durations as an exogenous factor contributing to the price discovery (the endogenous relationship between trades and prices), but also spreads and depths. We suggest that aside from the information content of incoming order flows (as suggested by Hasbrouck, 1991) order book illiquidity may also play a part in quote revisions.

Secondly, earlier works in market microstructure utilizing VAR models for estimating coefficients between quote revisions, trades and durations were conducted using ordinary least squares (OLS). Whilst OLS is acceptable with traditional empirical work in finance, market microstructure is unique in the sense that  $n$ , the sample size, is significantly larger. As  $n$  increases, the standard error shrinks considerably. This is trivial given the inverse relationship between the sample size and the standard error. Therefore, OLS is more susceptible to spurious relationships (more factors become significant), and hence a method with greater penalization is required. Therefore, we perform a subset selection procedure by adaptive lasso (Tibshorani, 1996; Zou, 2006). We are interested to see how this might impact the results documented in Dufour and Engle (2000).

Our results show that lagged spreads and depth plays a significant role in the formation of

quote revisions. Following our estimated VAR model, we construct price impact which accounts for order book illiquidity, and provides a more accurate indication of the information content of a trade (as per definition from Hasbrouck, 1991).

This topic is addressed in chapter 4.

### 1.3.3 Topic 3: Price impact's relationship with volatility

In extension to topic 2, we are able to construct time-varying price impact. If price impact varies significantly across time, then we are interested in its relationship with volatility. We attempt to answer this question from both an empirical and theoretical approach. Firstly, using the empirical results from topic 2, it is relatively straightforward to test its relationship with volatility. Aside from empirical testing, we describe a heterogeneous rational expectations model to theoretically motivate the relationship between price impact and volatility. We show higher price impact is related to higher overall volatility in the market. Price impact is higher in periods where there is higher information content (see Hasbrouck, 1991), so market makers and liquidity traders are more likely to react to incoming trades as there is a higher chance an informed trader is going to trade against them, this in turn increases volatility.

This topic is addressed in chapter 5.

### 1.3.4 Topic 4: Price impact, VPIN and the role of informed trader heterogeneity

Price impact measures the information content of incoming order flow and the risk aversion costs of being picked-off by an informed trader (see Escibano and Pascual, 2006), whilst VPIN measures order flow imbalance and the probability of being adversely selected (see Easley, Lopez de Prado and O'Hara, 2012). In this topic we examine the differences and similarities between these two measures. Initially, one might consider both measures to be similarly behaved. However,

## 1. INTRODUCTION

---

our empirical tests show that these two measures are negatively correlated. We conduct VAR modeling and show a contemporaneous negative relationship between price impact and VPIN, but no significant lag or lead. We study the empirical evidence and provide a theoretical model explaining the phenomenon. We show that if informed traders hold heterogeneous beliefs, then price impact would be negatively correlated with VPIN.

This topic is addressed in chapter 6.

Therefore, this thesis is structured as follows. Chapter 1 provides a general overview of material considered in this thesis and outlines several topics that will be examined. Chapter 2 provides an explanation of the empirical dataset, and methodology on data preparation. Chapter 3 addresses Topic 1, which introduces a new clustering methodology for examining information asymmetry and PIN. Chapter 4 addresses Topic 2, explains the price impact of a trade, and presents the econometric model used to estimate price impact. Chapter 5 addresses Topic 3, which examines the relationship between price impact and volatility. Chapter 6 addresses Topic 4 and compares price impact with VPIN.

# References

- [1] Diamond, D.W. and R.E. Verrecchia (1987) Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics* 18, 277 - 311
- [2] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [3] Easley, D., Hvidkjaer, S. and M. OHara, (2002) Is Information Risk a Determinant of Asset Returns?, *Journal of Finance* 10, 2185-2221
- [4] Easley, D., Hvidkjaer, S. and M. OHara, (2010) Factoring information into returns, *Journal of Financial and Quantitative Analysis* 45 - 2, 293 - 309
- [5] Easley, D., Kiefer, N., OHara, M., and J. Paperman, (1996) Liquidity, Information and Infrequently Traded Stocks, *Journal of Finance* 51, 1405-1436
- [6] Easley, D. and M. O'Hara (1992) Time and the process of security price adjustment, *Journal of Finance* 47, 577 - 605
- [7] Engle, R.F. and J. Lange (1997) Measuring, Forecasting and Explaining Time Varying Liquidity in the Stock Market, *NBER Working Papers 6129*, National Bureau of Economic Research, USA
- [8] Escribano, A. and R. Pascual (2006) Asymmetries in bid and ask responses to innovations in the trading process, *Empirical Economics* 30, 913 - 946
- [9] Escribano, A., Pascual., R. and M. Tapia (2004) Adverse selection costs, trading activity

## REFERENCES

---

- and price discovery in the NYSE: An empirical analysis *Journal of Banking & Finance* 28, 107 - 128
- [10] Glosten, L. and P. Milgrom (1985) Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71 - 100
- [11] Goldstein, M.A. and K.A. Kavajecz (2000) Eighths, sixteenth, and market depth: changes in tick size and liquidity provision on the NYSE, *Journal of Financial Economics* 56, 125 - 149
- [12] Hasbrouck, J. (1991) Measuring the information content of stock trades, *Journal of Finance* 46, 179 - 208
- [13] Hsu, N.J., Hung, H.L. and Y.M. Chang (2008) Subset selection for vector autoregressive process using Lasso, *Computational Statistics and Data Analysis* 52, 3645 - 3657
- [14] Kyle, A.S. (1985) Continuous auctions and insider trading, *Econometrica* 53 - 6, 1315 - 1336
- [15] Lin, W.W. and W.C. Ke (2011) A computing bias in estimating the probability of informed trading, *Journal of Financial Markets* 14, 625 - 640
- [16] Muranaga, J. (1999) Market microstructure and market liquidity, *IMES discussion paper series no.99 e.14*, Bank of Japan
- [17] Muranaga, J. and T. Shimizu (1999) Market microstructure and market liquidity, mimeo, Bank of Japan
- [18] O'Hara, M. (1995) *Market Microstructure Theory*, Blackwell, Cambridge, MA
- [19] Ren, Y.W. and X.S. Zhang (2010) Subset selection for vector autoregressive processes via adaptive Lasso, *Statistics and Probability Letters* 80, 1705 - 1712
- [20] Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B* 58-1, 267-288

## REFERENCES

---

- [21] Yan, Y. and S. Zhang (2012) An improved estimation method and empirical properties of the probability of informed trading, *Journal of Banking & Finance* 36, 454 - 467
- [22] Zou, H. (2006) The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101 No.476, 1418 - 1429

## REFERENCES

---

## 2

# Data

### 2.1 Tick Data

The empirical component of the research conducted in this dissertation is based on tick data from Thomson Reuters Tick History. Transactions data from Thomson Reuters allows us to collect incoming trades and best quote data. By doing so, we are able to easily construct the best bid and the best ask at any given point in time. Since the research we conduct do not extend beyond the first level of the orderbook, it is unnecessary to keep order book status (i.e., price and size at levels 2 and above). This bypasses the need to perform adaptive time window matching algorithms for synchronizing trade data and order book data as explained in Haustch and Huang (2012).

We remove the first 10 minutes of trading to eliminate abnormalities associated with the open, and remove any order book records where the spread between the best bid and ask could be negative or zero (this is also suggested in Haustch and Huang, 2012). In Dufour and Engle (2000) only the diurnal dummy for the start of the trading day is significant. Instead of parameterizing to account for opening anomalies, we simply remove them from our analysis. We justify this as the scope of the research is on examining trading behavior under normal circumstances rather than on the open.

Details on cleaning and synchronizing high frequency tick data are provided in the subsection 2.2.



## 2. DATA

---

Below we show a sample of the raw dataset we receive via Thomson Reuters Tick History for Hyundai Motors (005380.KS) for illustrative purposes.

There are 12 attributes in the dataset shown in Table 2.1. '#RIC' denotes the unique Reuters ticker code for a stock. '.KS' denotes the exchange code for the Korea Exchange, and '005380' denotes the unique ticker for Hyundai Motors. 'DateG' and 'TimeG' refers to the date and time in Greenwich Mean Time (GMT) to the microsecond<sup>1</sup>. 'GMTOffset' is used to convert GMT into local trading time. Here, the GMT offset is exactly 9 hours ahead. There are two classes within the Type attribute, namely trade and quote. Trade refers to a transaction taking place, and therefore to the right of a trade there will always be a price and volume entry. Trades can also be identified via the Qualifiers attribute. We note the 'NORMAL[GV5-TEXT]' entry denoted for trades. For the Korean exchange (similar to other Asian and North American exchanges), buy and sell initiations are not identified through the qualifier, therefore it is necessary to use an algorithm, such as the tick test or the Lee and Ready (1991) algorithm to determine initiation. This will be explained in section 2.3. Only the best bid and ask quotes are displayed in this dataset. Incoming quotes at the best bid and ask are displayed in the attributes 'BidSize' and 'AskSize'.

The raw dataset is not formatted in a manner that is easy for analysis. Firstly, we remove any anomalies using a method similar to a Bollinger band process described by Brownlees and Gallo (2006). In table 2.2 we show the snapshot of the actual dataset we use for analysis. Formatting is conducted in R<sup>2</sup>. For each transaction, we work out the date, time, price, volume, best bid price, best bid size, best ask price and best ask size. This completes the processing of trading attributes discussed in Chapter 1.

---

<sup>1</sup>The tables displayed here (tables 2.1 and 2.2) are only formatted to the second.

<sup>2</sup>Interested readers can contact the author for the R code used in this thesis. Email: weiwangchun@gmail.com

Table 2.1: Raw Transaction Data from Thomson Reuters Tick History

A generic snapshot of quote and trade data from the Thomson Reuters Tick History database. This particular snapshot is of Hyundai Motors, trading on the Korea Exchange (KRX). Date and time is in Greenwich Mean Time, with a GMT offset provided to calculate local time.

#RIC	DateG	TimeG	GMTOffset	Type	Price	Volume	BidPrice	BidSize	AskPrice	AskSize	Qualifiers
005380.KS	2/01/2007	1:00:41 AM		Trade	67700	4					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:00:41 AM		Trade	67700	5					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:00:41 AM		Trade	67700	46					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:00:41 AM		Quote			67600		67700		
005380.KS	2/01/2007	1:00:41 AM		Trade	67700	140				1814	NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:00:42 AM		Quote			972				
005380.KS	2/01/2007	1:00:55 AM		Trade	67700	300				1514	NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:00:56 AM		Quote							NORMAL[GV5_TEXT]:Low[USER]
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	1					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	9					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67700	100					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	91					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	876					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Quote			67500		67600		
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	8					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67600	25					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Trade	67700	13					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:01 AM		Quote					67700		
005380.KS	2/01/2007	1:01:01 AM		Quote					67600		
005380.KS	2/01/2007	1:01:04 AM		Quote				1063		10	
005380.KS	2/01/2007	1:01:09 AM		Trade	67600	10					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:09 AM		Trade	67700	40			67700		NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:09 AM		Quote							
005380.KS	2/01/2007	1:01:11 AM		Quote						1991	NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:29 AM		Trade	67700	55					
005380.KS	2/01/2007	1:01:29 AM		Quote			67600				
005380.KS	2/01/2007	1:01:29 AM		Trade	67700	100					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:29 AM		Trade	67600	73					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:29 AM		Trade	67600	3					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:29 AM		Quote				795		2091	NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:39 AM		Trade	67700	46					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:39 AM		Trade	67700	209					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:42 AM		Quote			1183			1886	
005380.KS	2/01/2007	1:01:50 AM		Trade	67600	10					NORMAL[GV5_TEXT]
005380.KS	2/01/2007	1:01:50 AM		Quote			1174			3156	

## 2. DATA

**Table 2.2:** Modified Transaction Data from Thomson Reuters Tick History

A generic snapshot of quote and trade data from the Thomson Reuters Tick History database. This particular snapshot is of Hyundai Motors, trading on the Korea Exchange (KRX). Date and time is in Greenwich Mean Time, while real trading time is +9 hours ahead. Each row represents a single transaction. Bid.Price, Bid.Size, Ask.Price and Ask.Size presents the status of the 1st level of the order book just prior to each transactions.

Date.G.	RTime	GMT.Offset	Price	Volume	Bid.Price	Bid.Size	Ask.Price	Ask.Size	Qualifiers
2/01/2007	1:05:47 AM	9	67600	1	67600	1937	67700	963	NORMAL[GV5.TEXT]
2/01/2007	1:05:57 AM	9	67700	20	67600	1936	67700	963	NORMAL[GV5.TEXT]
2/01/2007	1:05:58 AM	9	67700	250	67600	1936	67700	963	NORMAL[GV5.TEXT]
2/01/2007	1:06:03 AM	9	67700	69	67600	1946	67700	693	NORMAL[GV5.TEXT]
2/01/2007	1:06:16 AM	9	67700	16	67600	1961	67700	1624	NORMAL[GV5.TEXT]
2/01/2007	1:06:16 AM	9	67700	86	67600	1961	67700	1624	NORMAL[GV5.TEXT]
2/01/2007	1:06:16 AM	9	67700	152	67600	1961	67700	1624	NORMAL[GV5.TEXT]
2/01/2007	1:06:16 AM	9	67700	182	67600	1961	67700	1624	NORMAL[GV5.TEXT]
2/01/2007	1:06:16 AM	9	67700	61	67600	1961	67700	1624	NORMAL[GV5.TEXT]
2/01/2007	1:06:24 AM	9	67600	61	67600	1961	67700	1227	NORMAL[GV5.TEXT]
2/01/2007	1:06:24 AM	9	67600	286	67600	1961	67700	1227	NORMAL[GV5.TEXT]
2/01/2007	1:06:24 AM	9	67700	10	67600	1961	67700	1227	NORMAL[GV5.TEXT]
2/01/2007	1:06:31 AM	9	67700	19	67600	1414	67700	1217	NORMAL[GV5.TEXT]
2/01/2007	1:06:31 AM	9	67700	31	67600	1414	67700	1217	NORMAL[GV5.TEXT]

## 2.2 Data Preparation

We have chosen to use a methodology derived from Brownlees and Gallo (2006) for preparing high frequency data. It is used to clear out trade errors and data inconsistencies. This process is not dissimilar to Bollinger bands used in technical analysis. Essentially traded prices that are outside a rolling window threshold determined by a multiple of the sample standard deviation is eliminated. We have found this to be a simple yet effective measure for removing anomalies. Below we provide the steps,

Let  $p_0, p_1, p_2, \dots, p_I$  be the tick data series of traded price and  $v_0, v_1, v_2, \dots, v_I$  be the associated volume series for a particular day.

We decide to whether retain or remove a price volume pair  $(p_i, v_i)$  based on the following criteria,

$$\mathbf{1}_{\{\|p_i - \bar{p}_i(k)\| < 2\bar{s}_i(k) + \gamma\}} = \begin{cases} 1 : & (p_i, v_i) \text{ is kept} \\ 0 : & (p_i, v_i) \text{ is removed} \end{cases}$$

When  $\|p_i - \bar{p}_i(k)\|$  breaches the limit  $2\bar{s}_i(k) + \gamma$ , the price and volume  $(p_i, v_i)$  observation is removed.

$(\bar{p}_i(k), \bar{s}_i(k))$  denotes the sample mean and sample standard deviation of a neighborhood of  $k$  observations around sample  $i$ . More explicitly, we define it as the following, setting  $k = 2m + 1$

where integer  $m = 1, \dots, \lfloor \frac{I}{2} \rfloor$ .

1.  $(\bar{p}_i(k), \bar{s}_i(k)) = \left( \sum_{l=1}^k p_l, \frac{1}{k-1} \sum_{l=1}^k (p_l - \bar{p}_l(k))^2 \right)$  for  $i = 1, \dots, m$
2.  $(\bar{p}_i(k), \bar{s}_i(k)) = \left( \sum_{l=i-m}^{i+m} p_l, \frac{1}{k-1} \sum_{l=i-m}^{i+m} (p_l - \bar{p}_l(k))^2 \right)$  for  $i = m + 1, \dots, I - m$
3.  $(\bar{p}_i(k), \bar{s}_i(k)) = \left( \sum_{l=I-k+1}^I p_l, \frac{1}{k-1} \sum_{l=I-k+1}^I (p_l - \bar{p}_l(k))^2 \right)$  for  $i = I - m + 1, \dots, I$

The parameter  $\gamma$  is used to avoid cases where  $\bar{s}_i(k) = 0$  produced by a series of  $k$  equal prices. It is a fraction of traded price and is defined to be the minimum price variation allowed for the stock analyzed.

The adjustment of neighborhood parameter  $k$  or alternatively  $m$  and  $\gamma$  is crucial to cleaning the raw tick data and requires adjustment. In our data adjustment of stocks, we settle on using  $k = 41, \gamma = 0.0005 \cdot \bar{p}$  to produce the most satisfactory results. In frequently traded stocks, the size of the window  $k$  can be set reasonably large, whilst less liquid stocks might require a smaller value. We also note that in Brownlees and Gallo's (2006) paper,  $\gamma$  is a fixed value, rather than a function of average price as we have set here.

## 2.3 Trade Initiation

Trade initiation identifies whether the trade was initiated by the buyer or by the seller. Tick data from most North American and Asian exchanges do not have qualifier tags for determining trade initiation. Therefore several algorithms have been utilized by researchers to tackle this issue. For our research, we have used the Ellis, Michaely and O'Hara (2000) approach. To the best of our knowledge this has been the most recent trade initiation algorithm, and correctly classifies larger percentage of trades than the popular Lee and Ready (1991) approach. It has well been known that the tick rule is inaccurate. However, it is still popular with brokers where convenience outweighs accuracy. Below we describe each methodology.

## 2. DATA

---

### 2.3.1 Tick Rule

If the transaction price at time  $t$ ,  $p_t$ , is higher than  $p_{t-1}$ , then it is an *uptick* and would be classified as a buy initiation. If the transaction price  $p_t$  is lower than  $p_{t-1}$ , then it is a *downtick* and would be classified as a sell initiation. When there is no change in the last trade price at time  $t$ , then trade initiation follows the same classification as  $t - 1$ .

### 2.3.2 Lee and Ready (1991)

The Lee and Ready (1991) algorithm requires the best bid and ask prices. The mid price is defined to be the average between the best bid and the best ask. A five second lag between transaction price and the mid price is allowed, i.e., *the five second rule*. If the transaction price  $p_t$  is higher than the mid price, then it is classified as a buy initiation. If  $p_t$  is lower than the mid price, then it is classified as a sell initiation. If  $p_t$  is exactly on the mid price, then we revert to using the tick rule for classification.

### 2.3.3 Ellis, Michaely and O'Hara (2000)

All trades executed at the ask quote are classified as a buy initiation. All trades executed at the bid quote are classified as a sell initiation. All other trades are categorized by the tick rule. Ellis, Michaely and O'Hara (2000) compare their algorithm against Lee and Ready (1991) for NASDAQ stocks and show a higher rate of accuracy - 82.73% vs 80.77%. This is the approach, we implement in our thesis.

## 2.4 The Korea Exchange

The dataset we use for this research comes from the Korea Exchange. The Korea Exchange (KRX) is a continuous pure central limit order market with price/time priority with a total market capitalization of KRW 576,888,000 million. A single price call auction method is used for open and close. Circuit breakers exist on the KRX - stocks are restricted to a 15% change limit from previous closing prices. Trading is also suspended if there is a 10% fall (sustained for a minute) in the KOSPI index from the previous close. In 2010, the KRX changed its single fee

structure and introduced three separate fees, namely trading fee, settlement fee and access fee. KRX also lowered the tick size of of stocks less than KRW 1,000 from KRW 5 to KRW 1 on October 4th 2010. KRX recently introduced an auction-based block trading service to provide an anonymous liquidity pool. Orders are matched on continuous auction with time priority and are matched without quotes being displayed to the public. As of 2010, there was a total of 777 listed companies on the KRX.

### 2.4.1 Trading Rules

During continuous trading, any buy or sell order entered at a price that is equal to the ask or bid in the central limit order book will execute immediately. Once trades are executed, the volume will be deleted from the central limit order book. If the order volume cannot be executed completely, due to its size, the remaining volume enters the queue as a limit order. In instances where a market order is traded against several existing limit orders, the exchange generates a trade record for each market order - limit order pair of executing orders. In those instances, all multiple trade records generated by a single market order are aggregated into a single trade record. Trades and order prices are always visible to the public. Hidden orders (known as *icebergs*) will not impact transaction price and volume data, however it will impact quote sizes which subsequently impacts limit order analysis. Furthermore, Asian markets are also known for specific circuit breakers - Korean stocks are restricted to +/-15% change limit from previous closing prices. Korea also restricts naked short selling of financial stocks. Below we present the exchange timetable for Korea.

### 2.4.2 Data Sample

We pick 24 of the largest capitalization stocks on the Korea Exchange with complete tick data history from January 2007 to December 2012. These stocks are listed in table 2.4. The research conducted in the subsequent chapters will be based upon this dataset.

## 2. DATA

**Table 2.3:** Korea Daily Exchange Timetable

Korea Exchange (KRX)		
Market Stage	Time (local)	Functionality
Market Pre-open	8:00am to 8:59am	KOSPI: Orders may be entered, modified or deleted to join the opening auction at 9:00am
	7:50am to 8:59am	KOSDAQ: Orders may be entered, modified or deleted to join the opening auction at 9:00am
Pre-open off market crossings	7:30am to 8:30am	
Market trading	9:00am to 2:49pm	Continuous trading based on price/time.
Closing auction	2:50pm to 2:59pm	Orders may be entered, modified or deleted to join the closing auction at 3:00pm
Post market auction	3:00pm to 3:10pm	Orders may be entered, modified or deleted to trade at the closing price
Off hours closing price trading	3:10pm to 3:29pm	Shares can be traded at the closing price
Off hours single price trading	3:30pm to 6:00pm	Shares can be traded via auction every 30mins (+/- 5% from closing price)

*Source: Credit Suisse Global Markets and respective exchanges*

**Table 2.4:** Companies - Full List

We select the largest capitalization stocks listed on the Korea Exchange. However, we require stocks to have complete tick history data since at least January 2007. Below we list the 24 stocks we analyze in this paper. This information was obtained from the Korea Exchange website as at January 2013. Rank refers to the stock's ranking in the KOSPI 200. Market capitalization is in millions of Korean Won. Free-float rate is the percentage of the market capitalization that is freely available to be traded in the market. % KOSPI 200 refers to the percentage of the KOSPI that the stock makes up.

Rank	Reuters Ticker	Company	Market Cap.	% Free-float Rate	% KOSPI 200
1	005930 KS	Samsung Electronics	163,833,688	75	25.98
2	005380 KS	Hyundai Motors	29,065,481	70	4.61
3	005490 KS	POSCO	23,381,329	85	3.71
4	012330 KS	Hyundai Mobis	18,329,849	70	2.91
5	055550 KS	Shinhan Group	15,897,541	90	2.52
6	000660 KS	SK Hynix	15,493,666	80	2.46
7	000270 KS	Kia Motors	13,437,795	65	2.13
10	051910 KS	LG Chemicals	11,388,689	70	1.81
11	017670 KS	SK Telecom	10,654,397	70	1.69
13	015760 KS	KEPCO	10,335,622	50	1.64
15	009540 KS	Hyundai Heavy Industries	9,435,400	65	1.50
16	066570 KS	LG Electronics	9,147,913	65	1.45
17	033780 KS	KT&G	8,567,052	80	1.36
18	086790 KS	Hana Financial	8,127,584	95	1.29
19	030200 KS	KT Corp	8,001,119	85	1.27
20	000830 KS	Samsung C&T	7,834,321	85	1.24
22	034220 KS	LG Display	6,930,890	65	1.10
23	010950 KS	S-Oil	6,447,054	65	1.02
24	003550 KS	LG Corp.	5,922,161	55	0.94
25	010140 KS	Samsung Heavy Industries	5,541,009	75	0.88
26	051900 KS	LG Household & Healthcare	5,500,729	60	0.87
27	009150 KS	Samsung Electro-Mechanics	5,318,191	80	0.84
28	035250 KS	Kangwonland	4,269,183	65	0.68
29	000720 KS	Hyundai Engineering & Construction	4,198,112	65	0.67

# References

- [1] Brownlees, C.T. and G.M. Gallo (2006) Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics & Data Analysis* 51, 2232 - 2245
- [2] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [3] Ellis, K., Michaely R. and M. O'Hara (2000) The accuracy of trade classification rules: Evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* 35, 529-551.
- [4] Haustch, N. and R.H. Huang (2012) The market impact of a limit order, *Journal of Economic Dynamics and Control*, forthcoming
- [5] Lee, C. and M. Ready (1991) Inferring trade direction from intraday data, *Journal of Finance* 46, 733 - 746



## REFERENCES

---

### 3

# A Hierarchical Agglomerative Clustering approach for Estimating the Probability of Informed Trading

1

**Abstract:** *We present a new method for estimating the probability of informed trading. This method is called Cluster PIN (CPIN), and does not require maximum likelihood estimation of Poisson processes where floating point expectations may be an issue. Lin and Ke (2011) and Yan and Zhang (2012) show Easley, Hvidkjaer and O'Hara's (2002, 2010) PIN is biased and underestimates the true level for large cap stocks; they provide a solution. We show that CPIN which is simple and robust under different trading circumstances, and is 300x faster to compute than Lin and Ke (2011) and Yan and Zhang's (2012) remedy and comparable in terms of accuracy. We suggest that researchers can either use CPIN directly or input it as the starting value of Lin and Ke's (2011) method. Furthermore, CPIN is able to provide researchers with explicit classification of the status of trading (good news, bad news, no news) on a daily basis, this cannot be achieved with MLE PIN, and allows us to study the behavior of trades around information days.*

---

<sup>1</sup>The R code for simulation and estimation of CPIN can be provided upon request. It utilizes the R package 'hclust'.

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

#### 3.1 Introduction

This study introduces an alternative methodology for estimating the probability of informed trading (PIN) (see Easley, Kiefer, O’Hara and Paperman, 1996; Easley, Hvidkjaer and O’Hara, 2002; and Easley, Hvidkjaer and O’Hara, 2010). We show our new method is capable of performing PIN estimation at a fraction of the speed and also possess the ability to explicitly classify days into good news, bad news and no news; this ability is useful for understanding trading behavior and also allows us to perform ex-post analysis and validations on the classifications.

PIN is a widely used variable in market microstructure for detecting the level of information asymmetry or informed trading. PIN assumes news events are drawn on a daily basis from a Bernoulli random variable  $Bin(1, p)$ , with a constant parameter  $p$ . Likewise the conditional event of bad news for a particular day is drawn from another constant parameter Bernoulli random variable. Buy and sell initiated trades are assumed to be transacted with exponential waiting time; therefore for a given time interval the aggregate number of buys or sells can be modeled using Poisson distributions. Two Poisson processes are required, one for buys and one for sells. The intensity of each process is determined by the presence of informed and uninformed traders. For instance, informed traders will only be buying if there is good news and selling if there is bad news. In essence, the PIN model is a Bernoulli modulated Poisson process and the PIN metric itself relates to the expected percentage of informed trading.

In Easley, Hvidkjaer and O’Hara (2010) it was noted that a large number of buys and sells might be problematic in the maximum likelihood procedure. They perform log-likelihood factorization (reorganizing the log-likelihood expression) to reduce this problem by making it more computationally efficient. Recently, Lin and Ke (2011) and Yan and Zhang (2012) show further computational bias in the existing literature on PIN. Lin and Ke (2011) identifies a computing bias due to floating point exceptions (FPE) which is particularly evident for active stocks. FPE may interfere with R, Matlab or SAS in finding the optimal solution. Furthermore, they show that Easley, Hvidkjaer and O’Hara’s (2010) likelihood factorization technique is computationally inaccurate and leads to a downwards bias. They offer a more accurate likelihood expression,

which they prove using simulations. Yan and Zhang (2012) develop an initial values algorithm to avoid boundary solutions and reduce the chance of obtaining a local maxima.

These two papers have provided major improvement in PIN estimation over the original methods proposed in Easley, Kiefer, O'Hara and Paperman (1996) and Easley, Hvidkjaer and O'Hara (2002, 2010). It has, therefore, been recommended that empirical microstructure researchers use Lin and Ke (2011) and Yan and Zhang (2012) methods to re-estimate their PIN values. However, one main weakness with Lin and Ke (2011) and Yan and Zhang (2012) is that estimation speed is relatively slow. The initial value algorithm requires users to loop through 125 initial value combinations and work out all their log likelihoods. If this exhaustive initial value algorithm is ignored, PIN results will be inaccurate as the MLE might have only yielded a local maxima. On our computers (Dell Optiplex 980 Intel i5 650 @ 3.19 GHz) the time taken to estimate one PIN value on 100 days worth of buy / sell values took as long as 14 seconds to complete. If one had to estimate PIN values across time for 10 years or cross-sectionally for 2000 stocks, this procedure would slow dramatically to several hours or days.

This paper introduces a new method for estimating PIN. We show our new method provides accuracy comparable to that of Lin and Ke (2011) and Yan and Zhang (2012) (and far superior to that of Easley, Hvidkjaer and O'Hara, 2002, 2010), yet computed at less than a fraction of the time. On the same computer, using the same dataset as the previous example, our method took only 0.037 seconds to estimate PIN. We call our method cluster PIN (CPIN). It uses hierarchical agglomerative clustering (HAC) algorithms to classify days into three groups: days with good news, days with bad news and days with no news. By knowing exactly which days have been classified in which group, we can work out the frequency of news and approximate the intensity of informed and uninformed traders, knowing that the intensity of a Poisson distribution can be simply estimated from the average of its realizations. We show PIN values estimated via clusters to be fairly accurate and robust. Its biggest advantage is that it bypasses FPE completely and estimates at a magnitude of 300x faster than Lin and Ke (2011) and Yan and Zhang's (2012) method.

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

Our first objective is to compare the accuracy, speed and robustness of the following techniques:

1. YZ-EHO-PIN: Easley, Hvidkjaer and O'Hara's (2010) methodology on the log-likelihood with Yan and Zhang's (2012) algorithm for initial values
2. YZ-LK-PIN: Lin and Ke's (2011) methodology on the likelihood with Yan and Zhang's (2012) algorithm for initial values
3. CPIN: Our cluster PIN methodology using three different linkage methods: complete-linkage clustering (farthest neighbor), average-linkage clustering (UPGMA) and Ward clustering (minimum variance criterion)

We also show that our method compliments that of Lin and Ke (2011). Users can bypass the exhaustive initial value algorithm suggested in Yan and Zhang (2006, 2012) by using CPIN first to determine an initial set of parameters for Lin and Ke's (2011) MLE. We show that the CPIN estimates are extremely accurate to begin with, and so MLE is unlikely to yield a local maxima. Our results show that Lin and Ke's (2011) MLE together with CPIN initial values to be more accurate and speedier than Lin and Ke's (2011) MLE with Yan and Zhang's (2006, 2012) initial value algorithm.

Our second objective is to illustrate the benefits of the CPIN methodology in being able to explicitly classify days into good news days, bad news days and no news days. We conduct an ex-post examination of the differences in trading behavior between these three clusters. We find daily returns were higher in the good news group when compared to the bad news group; this confirms our CPIN classification is sensible. Likewise, we found both good news days and bad news days had higher realized volatility than no news days; this is consistent with findings on a positive relationship between volatility and information (Clark, 1973; Ross, 1989). It was also not surprising that buy initiated volume was larger in good news days and sell initiated volume was larger in bad news days. Finally, we also show that the classification of news holds autoregressive properties. This points to news being 'sticky', and information dissemination

may take several days to complete.

The paper is structured as follows. Section 3.2 provides a brief discussion on existing PIN methodologies. Section 3.3 introduces our new method CPIN. We run a horse race between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations in section 3.4. Section 3.5 illustrates how to incorporate CPIN with the existing Lin and Ke method and section 3.6 analyzes trading characteristics between good and bad news days which are classified by CPIN. We conclude in section 3.7.

## 3.2 Estimating PIN

We first explain Easley, Hvidkjaer and O'Hara's (2002) numerical estimation method for PIN, as it is the most widely used. As discussed by Lin and Ke (2011), this method suffers from computing bias due to the effect of the floating-point exception. Therefore, we follow by discussing the modified approach of Lin and Ke (2011). Furthermore, we discuss a new starting value algorithm described in Yan and Zhang (2012) and utilize it for both approaches.

The joint probability density function for the observed number of buys  $B_t$  and sells  $S_t$  on day  $t$  is,

$$\begin{aligned}
 f(B_t, S_t | \theta) = & \alpha \delta e^{-\varepsilon_b} \frac{\varepsilon_b^{B_t}}{B_t!} e^{-(\varepsilon_s + \mu)} \frac{(\varepsilon_s + \mu)^{S_t}}{S_t!} \\
 & + \alpha(1 - \delta) e^{-(\varepsilon_b + \mu)} \frac{(\varepsilon_b + \mu)^{B_t}}{B_t!} e^{-\varepsilon_s} \frac{\varepsilon_s^{S_t}}{S_t!} \\
 & + (1 - \alpha) e^{-\varepsilon_b} \frac{\varepsilon_b^{B_t}}{B_t!} e^{-\varepsilon_s} \frac{\varepsilon_s^{S_t}}{S_t!}
 \end{aligned} \tag{3.1}$$

where  $\alpha$  is the probability of a news or information event,  $\delta$  is the conditional probability of bad news,  $\mu$  is the informed trader intensity and  $\varepsilon_b$  and  $\varepsilon_s$  are arrival rates of uninformed buy and sell trades. PIN assumes independent news events which occur at daily frequency. The log-likelihood is simply  $L(\theta|B, S) = \sum_{t=1}^T \log f(B_t, S_t | \theta)$ . To estimate  $\theta$ , we solve  $\arg \max_{\theta} L(\theta|B, S)$ . As shown in Easley, Hvidkjaer and O'Hara's (2002), after obtaining  $\hat{\theta}$ , PIN is readily estimated as  $P\hat{I}N = \frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu} + \hat{\varepsilon}_b + \hat{\varepsilon}_s}$ .

Easley, Hvidkjaer and O'Hara's (2002) note that when performing maximum likelihood es-

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

timization (MLE) yielded 716 non-convergence and 456 cases of corner solutions from approximately 20,000 stocks. Brown, Hillegeist and Lo (2004) filtered between 14 - 19% of the sample due to corner solutions. Two main reasons are blamed for optimization failure. (1) Optimization algorithms locate only local maxima/minima rather than global maxima/minima. (2) Computational overflow and underflow occurs due to the floating point exception (FPE). Most high-level languages used by researchers, associate  $e^{708}$  and  $e^{-708}$  (SAS),  $e^{710}$  and  $e^{-745}$  (Matlab),  $e^{710}$  and  $e^{-746}$  (R) with overflow and underflow respectively.

Let us discuss problem (2) first. It is trivial that in Easley, Hvidkjaer and O'Hara's (2002), forms such as  $e^{\varepsilon_b + \mu}$  is likely to cause floating point exceptions. To reduce this problem, Easley, Hvidkjaer and O'Hara (2010) consider log-likelihood factorization. This is done to achieve a degree of computational efficiency and reduce 'truncation error' due to FPE. The daily log joint density function in Easley, Hvidkjaer and O'Hara (2010) is,

$$\begin{aligned} \log f(B_t, S_t | \theta) = & \log[\alpha \delta e^{-\mu} x_b^{B_t - M_t} x_s^{-M_t} + \alpha(1 - \delta) e^{-\mu} x_b^{-M_t} x_s^{S_t - M_t} + (1 - \alpha) x_b^{B_t - M_t} x_s^{S_t - M_t}] \\ & + B_t \log(\varepsilon_b + \mu) + S_t \log(\varepsilon_s + \mu) - (\varepsilon_b + \varepsilon_s) + M_t [\log(x_b) + \log(x_s)] - \log(B_t! S_t!) \end{aligned} \quad (3.2)$$

where  $M_t = \min(B_t, S_t) + \max(B_t, S_t)/2$ ,  $x_b = \varepsilon_b / (\mu + \varepsilon_b)$  and  $x_s = \varepsilon_s / (\mu + \varepsilon_s)$ . Immediately it is obvious that the last term,  $\log(B_t! S_t!)$ , is a constant and can be dropped out.

When estimating  $\theta = \{\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s\}$  we consider the set of basic feasible solutions (BFS) where the lower boundary is  $\{0, 0, 0, 0, 0\}$  and the upper boundary is  $\{1, 1, Inf, Inf, Inf\}$ . Lin and Ke (2011) explains that the actual BFS we use to maximize our log-likelihood is smaller than the theoretical BFS. It is smaller because instances where  $\theta$  leads to FPE are removed from the set. Therefore, they show FPE leads to selection bias in Easley, Hvidkjaer and O'Hara (2002).

Furthermore, Lin and Ke (2011) argue Easley, Hvidkjaer and O'Hara's (2010) likelihood expression after factorization is inaccurate for computation and has an inherent downwards bias. They argue their approach to remedy FPE provides a more accurate likelihood expression.

The daily log joint density function in Lin and Ke (2011) is,

$$\begin{aligned} \log f(B_t, S_t | \theta) = & \log[\alpha \delta e^{(k_{1,t} - k_{max,t})} + \alpha(1 - \delta)e^{(k_{2,t} - k_{max,t})} + (1 - \alpha)e^{(k_{3,t} - k_{max,t})}] \\ & B_t \log(\varepsilon_b + \mu) + S_t \log(\varepsilon_s + \mu) - (\varepsilon_b + \varepsilon_s) + k_{max,t} - \log(B_t! S_t!) \end{aligned} \quad (3.3)$$

where  $k_{1,t} = -\mu - B_t \log(1 + \mu/\varepsilon_b)$ ,  $k_{2,t} = -\mu - S_t \log(1 + \mu/\varepsilon_s)$ ,  $k_{3,t} = -B_t \log(1 + \mu/\varepsilon_b) - S_t \log(1 + \mu/\varepsilon_s)$ , and  $k_{max,t} = \max(k_{1,t}, k_{2,t}, k_{3,t})$ . Similarly, the constant term,  $\log(B_t! S_t!)$ , is unnecessary. Their approach is derived from 2 computing principles. Firstly, they note that computing the expression  $e^{x+y}$  is more stable than computing  $e^x e^y$ . Secondly, a large input for  $\exp(\cdot)$  or a small input for  $\log(\cdot)$  should be avoided.

In Lin and Ke (2011), initial values are chosen using the Yan and Zhang (2006, 2012) method. This deals with the local maxima/minima problem we mentioned earlier. In our paper, we will utilize this approach for both Easley, Hvidkjaer and O'Hara's (2010) and Lin and Ke's (2011) likelihood expressions. The initial values  $\theta^0 = (\alpha^0, \delta^0, \mu^0, \varepsilon_b^0, \varepsilon_s^0)$  are,

$$\begin{aligned} \alpha^0 = \alpha_i, \delta^0 = \delta_j, \varepsilon_b^0 = \gamma_k \bar{B}, \mu^0 = (\bar{B} - \varepsilon_b^0)/(\alpha^0(1 - \delta^0)), \text{ and } \varepsilon_s^0 = \bar{S} - \alpha^0 \delta^0 \mu^0, \text{ if } \bar{B} \leq \bar{S} \\ \alpha^0 = \alpha_i, \delta^0 = \delta_j, \varepsilon_s^0 = \gamma_k \bar{S}, \mu^0 = (\bar{S} - \varepsilon_s^0)/(\alpha^0 \delta^0), \text{ and } \varepsilon_b^0 = \bar{B} - \alpha^0(1 - \delta^0)\mu^0, \text{ if } \bar{B} > \bar{S} \end{aligned}$$

where  $\bar{B}$  and  $\bar{S}$  are the sample means for the number of buy and sell trades respectively on a daily frequency. Yan and Zhang (2012) construct 125 initial value combinations where  $\alpha_i, \delta_j$  and  $\gamma_k$  take their values from the set  $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ . Combinations resulting in negative arrival intensities are removed. MLE is performed on all the remaining combinations, non-convergence cases are removed and the solution with the maximum log likelihood is chosen. Whilst this approach ensures a global maxima is determined, it is approximately 125x slower than single iteration MLE.

### 3.3 Cluster PIN

In this study we consider a new methodology to estimate PIN that bypasses MLE of joint Poisson processes completely. We show that whilst our methodology is radically different, the results are the same as MLE PIN and 300x faster to compute. Cluster PIN (CPIN) is based



### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

on hierarchical agglomerative clustering (HAC) techniques. These techniques are generally used for classification, i.e., to cluster or group elements based on a distance matrix. Rarely would one think of applying it to time-series data, but here we consider clustering days into 3 clusters. Cluster 1 consists of days where there is good news with informed intensity on the buys. Cluster 2 consists of days where there is bad news with informed intensity on the sells. Cluster 3 consists of days where there is no news, and hence only uninformed intensity on both sides of the book. After the construction of the 3 clusters, it is not difficult to approximate informed and uninformed intensities. The benefit of forming clusters is that we are able to associate an individual day precisely to a cluster and point out whether it was a good news day, a bad news day or a no news day. This type of granularity cannot be achieved with PIN MLE methods.

HAC initially treats each element (in our case each day) as a cluster of its own (i.e., singletons). The clusters are sequentially merged into larger clusters, until all elements are grouped into one single cluster. This can be expressed in a dendrogram. In our application, the elements are in a one-dimensional space and the clusters sequentially merge and stop when we have three clusters. At each stage, HAC combines the two *nearest* clusters into one. To determine what is *nearest*, a distance measure is required. HAC variations are caused by different definition of cluster distances. Here we consider three different HAC techniques: complete, average and Ward. A commonly used HAC method, single-linkage or nearest neighbor clustering is not used, as it is susceptible to noise and outliers which is common with financial data.

1. 'Complete' refers to complete-linkage clustering or farthest neighbor clustering (Defays, 1977). In the beginning of the process, each element is in a cluster of its own. The clusters are then sequentially combined into larger clusters, until all elements end up being in the same cluster. At each step, the two clusters separated by the shortest distance are combined. In complete-linkage clustering, the link between two clusters contains all element pairs, and the distance between clusters equals the distance between those two elements (one in each cluster) that are farthest away from each other. The shortest of these links that remains at any step causes the fusion of the two clusters whose elements

are involved. Let  $X$  and  $Y$  be two clusters and  $D(X, Y)$  be the distance between them,

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$$

2. 'Average' refers to average-linkage clustering or UPGMA (unweighted pair group method with arithmetic mean) see Sokal and Michener (1958). The distance between any two clusters  $X$  and  $Y$  is taken to be the average of all distances between pairs of objects "x" in  $X$  and "y" in  $Y$ , that is, the mean distance between elements of each cluster:

$$D(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)$$

3. 'Ward' refers to Ward's method, see Ward (1963). Ward suggested a general agglomerative hierarchical clustering procedure, where the criterion for choosing the pair of clusters to merge at each step is based on the optimal value of an objective function. To illustrate the procedure, Ward used the example where the objective function is the error sum of squares, and this example is known as Ward's method or more precisely Ward's minimum variance method. Ward's minimum variance criterion minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers. At the initial step, all clusters are singletons.

Below we provide our method and also provide a simple example involving 94 daily observations from Hyundai Motors (005380 KS) from 02-January 2007 to 16-May 2007.

1. We construct a net order flow imbalance series,  $X_t = B_t - S_t \forall t = 1$  to  $T$ , where  $B_t$  is the number of buy initiated trades in day  $t$  and  $S_t$  is the number of sell initiated trades in day  $t$ .

*Example:* We extract tick data from 02-January 2007 to 16-May 2007 on trades for Hyundai Motors. We determine buy or sell initiated trades through the implementation of

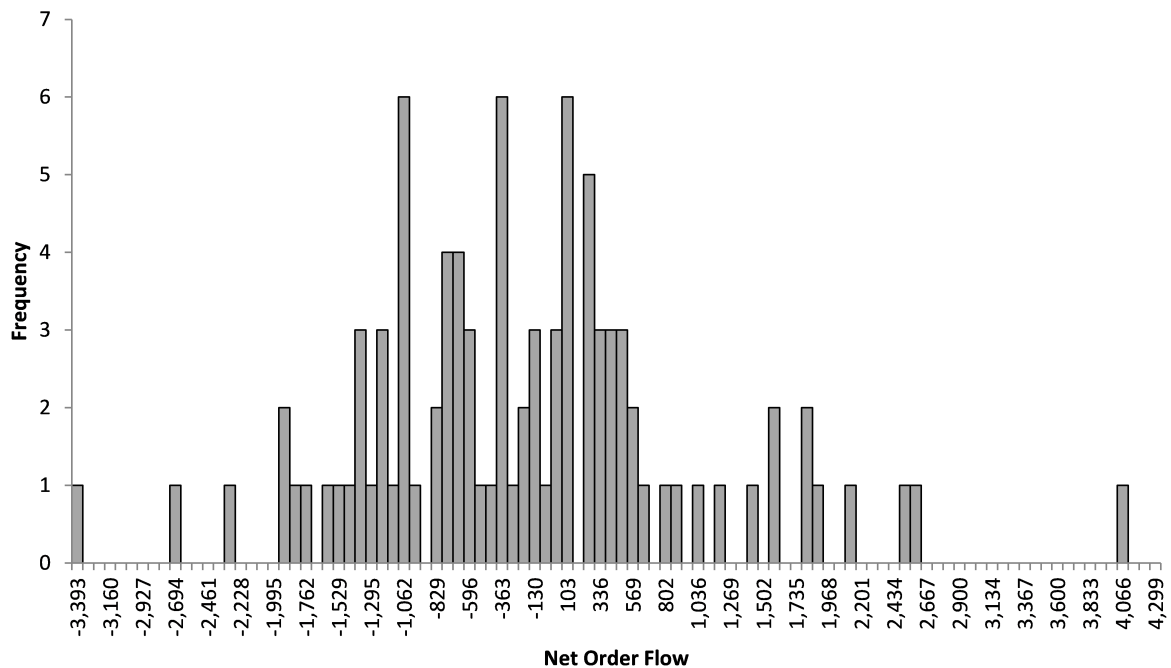
### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

*Ellis, Michaely and O'Hara's (2000) algorithm. Following this, we can determine daily  $B_t$  and  $S_t$ . In figure 3.1 we show the histogram of the net order flow for Hyundai Motors.*

**Figure 3.1:** Net Order Flow Imbalance Histogram

The distribution of daily net order flow imbalance  $X_t$  for Hyundai Motors over 94 days from 2nd January 2007 to 16th May 2007.



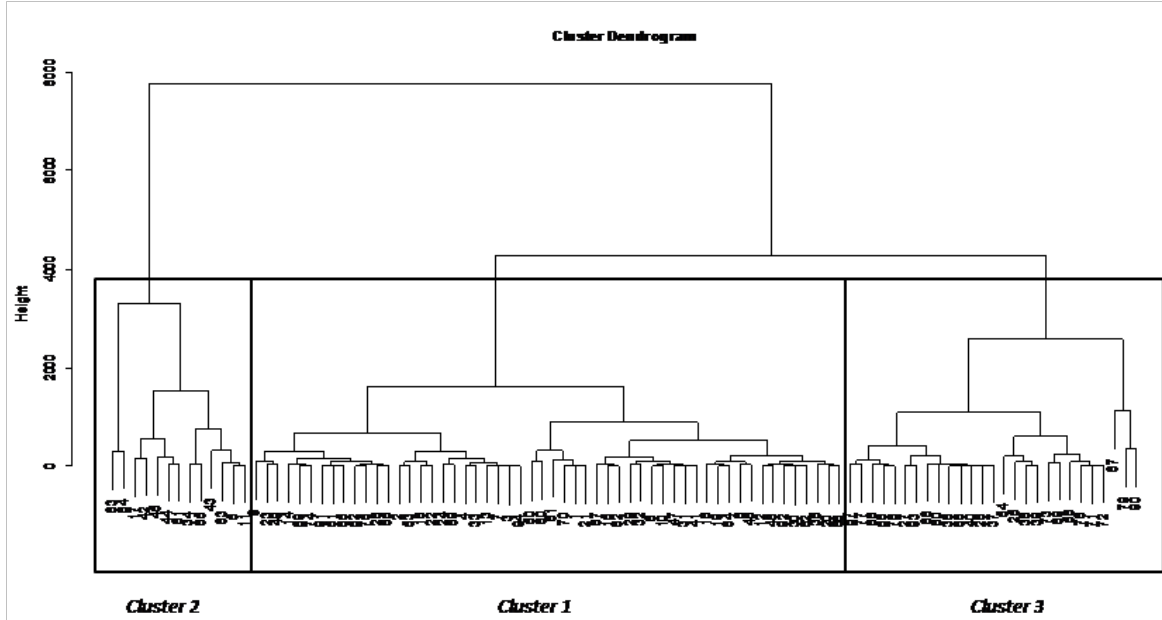
2. Single dimension HAC is performed on  $X_t$  using one of three distance methods (complete, average or Ward). It is single dimension because the distance measure is based on only one factor,  $X_t$ . The hierarchical cluster dendrogram is then cut into exactly 3 clusters. each cluster contains a number of days  $t$ .

*Example: In figure 3.2 we show the dendrogram for Hyundai Motors. The numbers at the nodes represent the days index  $t$ . From the dendrogram we can see HAC groups individual days to ultimately one single cluster, however we stop when it reaches 3 clusters.*

3. The cluster with the largest average  $X_t$  is defined as the good news cluster ( $\uparrow$ ). Likewise the cluster with the smallest average  $X_t$  is defined as the bad news cluster ( $\downarrow$ ). The remaining cluster is the no news cluster ( $-$ ). For each cluster, we calculate the mean daily

**Figure 3.2:** Net Order Flow Imbalance Dendrogram

The dendrogram from daily net order flow imbalance  $X_t$  for Hyundai Motors over 94 days from 2nd January 2007 to 16th May 2007.



$B_t$  and  $S_t$  which we denote  $\bar{B}_c$  and  $\bar{S}_c$  for  $c \in \{\uparrow, \downarrow, -\}$ . We also calculate the percentage time each cluster occupies of the total time  $T$  as  $\omega_c$  for  $c \in \{\uparrow, \downarrow, -\}$  such that  $\sum_c \omega_c = 1$ . A simple table such as the one below can be populated,

	Good News $\uparrow$	Bad News $\downarrow$	No News $-$
buys	$\bar{B}_\uparrow$	$\bar{B}_\downarrow$	$\bar{B}_-$
sells	$\bar{S}_\uparrow$	$\bar{S}_\downarrow$	$\bar{S}_-$
weight	$\omega_\uparrow$	$\omega_\downarrow$	$\omega_-$

*Example:* We populate the table for Hyundai Motors using data from 2nd January 2007 to 16th May 2007.

	Good News $\uparrow$	Bad News $\downarrow$	No News $-$
buys	$\bar{B}_\uparrow = 4349.4$	$\bar{B}_\downarrow = 2226.0$	$\bar{B}_- = 2144.3$
sells	$\bar{S}_\uparrow = 2248.4$	$\bar{S}_\downarrow = 3731.2$	$\bar{S}_- = 2231.0$
weight	$\omega_\uparrow = 13/94$	$\omega_\downarrow = 27/94$	$\omega_- = 54/94$

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

4. We approximate the arrival intensities and probabilities, the underlying components of PIN. Firstly, we work out the informed buy and sell intensities. We match the first moment<sup>1</sup> and the informed buy and sell intensities are:

(a) Informed buy intensity  $\hat{\mu}_b = \bar{B}_\uparrow - \bar{S}_\uparrow$

(b) Informed sell intensity  $\hat{\mu}_s = \bar{S}_\downarrow - \bar{B}_\downarrow$

We then estimate the 5 underlying PIN components as described in Easley, Hvidkjaer and O'Hara (2002,2010).

(a) Informed intensity  $\hat{\mu} = \frac{\omega_\uparrow}{\omega_\downarrow + \omega_\uparrow} \hat{\mu}_b + \frac{\omega_\downarrow}{\omega_\downarrow + \omega_\uparrow} \hat{\mu}_s$

(b) Uninformed buy intensity  $\hat{\varepsilon}_b = \omega_\uparrow(\bar{B}_\uparrow - \hat{\mu}_b) + \omega_\downarrow \bar{B}_\downarrow + \omega_- \bar{B}_-$

(c) Uninformed sell intensity  $\hat{\varepsilon}_s = \omega_\downarrow(\bar{S}_\downarrow - \hat{\mu}_s) + \omega_\uparrow \bar{S}_\uparrow + \omega_- \bar{S}_-$

(d) Probability of news  $\hat{\alpha} = \omega_\downarrow + \omega_\uparrow$

(e) Conditional probability of bad news  $\hat{\delta} = \omega_\downarrow / \hat{\alpha}$

*Example: We work out the arrival intensities and probabilities for Hyundai Motors.*

(a) Informed buy intensity  $\hat{\mu}_b = 2101$

(b) Informed sell intensity  $\hat{\mu}_s = 1505.2$

(c) Informed intensity  $\hat{\mu} = 1698.8$

(d) Uninformed buy intensity  $\hat{\varepsilon}_b = 2182.2$

(e) Uninformed sell intensity  $\hat{\varepsilon}_s = 2232.0$

(f) Probability of news  $\hat{\alpha} = 40/94$

(g) Conditional probability of bad news  $\hat{\delta} = 27/40$

5. Finally, CPIN is estimated as  $\frac{\hat{\alpha}\hat{\mu}}{\hat{\alpha}\hat{\mu} + \hat{\varepsilon}_b + \hat{\varepsilon}_s}$

*Example: CPIN for Hyundai Motors is 0.141. To check, PIN for the same dataset using Lin and Ke's (2011) method and Yan and Zhang's (2012) initial value algorithm produces 0.142, where  $\hat{\alpha} = 0.34$ ,  $\hat{\delta} = 0.53$ ,  $\hat{\mu} = 2141.34$ ,  $\hat{\varepsilon}_b = 2117.06$  and  $\hat{\varepsilon}_s = 2291.06$ .*

---

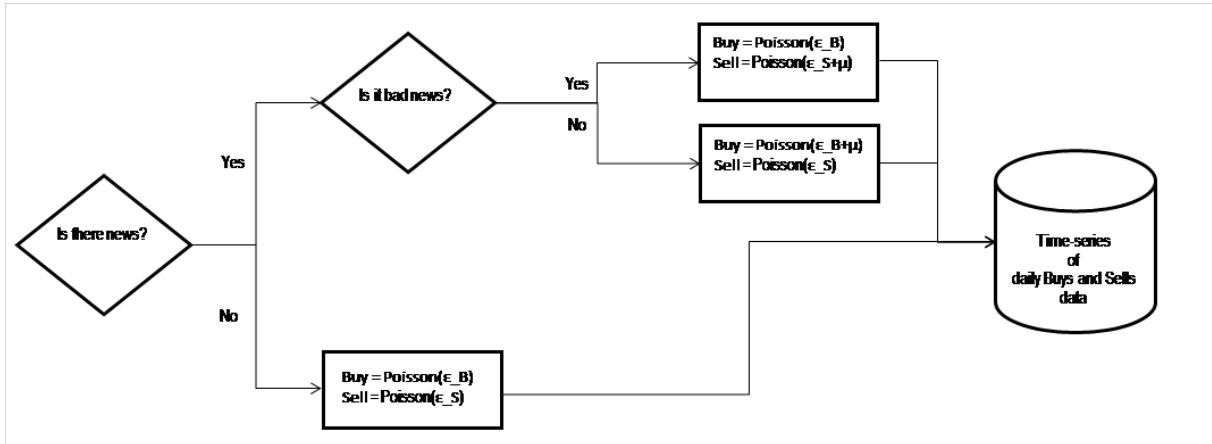
<sup>1</sup>The intensity  $\lambda$  parameter in a Poisson distribution ( $Z \sim Poi(\lambda)$ ) is estimated as the average  $\bar{z}$ , since  $E(Z) = \lambda$

### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

Here we simulate buy and sell time-series using Easley, Hvidkjaer and O’Hara’s (2002) PIN model and test whether the estimates provided by the methods YZ-EHO, YZ-LK and clusters are accurate.

Buy and sells are simulated using two Bernoulli distributions (news and bad news). If the news random variable yields a negative realization then the number of buys and sells for that day would be generated by  $Poi(\varepsilon_B)$  and  $Poi(\varepsilon_S)$  respectively. If the news random variable yields a positive realization and the bad news random variable generates a negative realization, then the number of buys and sells for that day would be generated by  $Poi(\varepsilon_B + \mu)$  and  $Poi(\varepsilon_S)$  respectively (and  $Poi(\varepsilon_B)$  and  $Poi(\varepsilon_S + \mu)$  if the bad news random variable generated a positive realization).

Figure 3.3: Buy and Sell Generation



In our first set of simulations we test the computing bias between estimation methods. Lin and Ke (2011) criticize Easley, Hvidkjaer and O’Hara’s (2010) method to underestimate actual PIN when the number of daily trades is high. Here we test the question: are PIN estimates dependent on the size of daily number of buys and sells? We set the theoretical PIN value to

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

be  $PIN = 0.1111$  ( $\alpha = \frac{1}{2}$ ,  $\delta = \frac{1}{2}$ ,  $\mu = 0.2k$ ,  $\varepsilon_B = 0.4k$ ,  $\varepsilon_S = 0.4k$ ) and we vary  $k$  to increase the number of trades. We test  $k = 50$  to 5000 at steps of 50, i.e., 50, 100, 150, ... etc. At each  $k$ , we generate 50 sets of buy and sell time-series. Each time-series consists of 100 days worth of observations. For each time series we estimate PIN and PIN parameters using five methods (YZ-EHO-PIN, YZ-LK-PIN, CPIN complete-linkage, CPIN average-linkage, CPIN Ward's method).

Figure 3.4 shows significant downwards bias using YZ-EHO-PIN when the daily number of trades exceed 2000. This is consistent with findings in Lin and Ke (2011). However, we do not find any major bias with YZ-LK-PIN or CPIN measures. CPIN using average-linkage and Ward's method show some bias when the total number of trades is low. Figure 3.5 shows that variance of PIN estimates is higher when the number of trades per day is lower. This suggests lower accuracy and higher variability for smaller capitalization stocks. YZ-LK-PIN, CPIN (complete linkage) and CPIN (Ward's method) show similar levels of variability whilst YZ-EHO-PIN and CPIN (average linkage) are somewhat higher. Our results suggest YZ-LK-PIN and CPIN (complete linkage) are the superior methods from the five and comparable in performance between each other. Figure 3.6 shows the computing time it takes to estimate PIN. It is clear that CPIN methods are much faster, taking on average 0.037 seconds to compute a PIN value. YZ-EHO-PIN took as long as 22 seconds to compute. The gradual reduction is due to the failure to converge when number of trades are high. YZ-LK-PIN took an increasingly longer time to compute as the number of trades went up, from approximately 11 seconds to 14 seconds. This makes cluster PIN approximately 300x faster to run.

We are not particularly surprised at the overwhelming difference in speed between our method and existing methods. In our method there is no estimation of latent parameters, we are simply classifying trading days into clusters. The time complexity of HAC methods are known to have quadratic time  $O(n^2)$ ; this is because at the first iteration all HAC methods need to compute the distance of all pairs of the  $n$  elements. Methods discussed in Easley, Hvidkjaer and O'Hara (2010), Lin and Ke (2011) and Yan and Zhang (2012) use MLE which performs

### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

---

convex optimization. It is known that the time complexity of MLE is polynomial time  $O(2^n)$  which is considerably slower.

Our second set of simulations also concerns with the comparison of actuals versus estimates. Here we run 1000 simulated buy and sell time-series of 100 days in length each from randomly generated PIN parameters. We let  $\alpha$  - P(news),  $\delta$  - P(bad—news),  $\mu$  - informed intensity,  $\varepsilon_B$  - uninformed buy intensity and  $\varepsilon_S$  - uninformed sell intensity to be structured as follows:

$$\alpha = a; \quad \delta = b; \quad \mu = ck; \quad \varepsilon_B = \frac{1-c}{2}k; \quad \varepsilon_S = \frac{1-c}{2}k$$

$k$  resembles total trade intensity (we set it at 2,500). Given the Poisson definition,  $k$  also is the average daily number of trades. We randomize by giving setting  $a$ ,  $b$  and  $c$  to be random variables  $\sim U[0, 1]$ .

We then estimate the PIN from the buy and sell data using three methods: (1) YZ-EHO-PIN (2) YZ-LK-PIN and (3) CPIN (complete-linkage) (We have already determined that complete-linkage to be superior than Ward and average from our first set of test). Further from just comparing actual and estimated PIN values, we also compare actual and estimated PIN parameters. In Lin and Ke (2011) it was shown that informed intensity  $\mu$  is generally underestimated and uninformed intensity  $\varepsilon_B$  and  $\varepsilon_S$  is overestimated when one uses Easley, Hvidkjaer and O'Hara's (2005, 2010) method. Here we test the accuracy of parameter estimation between existing methods and our new cluster method.

Figure 3.7 illustrates the accuracy of YZ-EHO-PIN. It is obvious that YZ-EHO-PIN grossly underestimates actuals when PIN is above 0.2. Consistent with Lin and Ke's (2011) findings, we find informed intensity is underestimated whilst uninformed intensity is overestimated. Whilst there is no obvious bias with estimating the probability of news and the conditional probability of bad news, we can see that it is not particularly accurate. Figure 3.8 illustrates the accuracy of YZ-LK-PIN. It is significantly better than YZ-EHO-PIN. Firstly, we note no bias in estimation when actual PIN are high. Also, there are no biases for informed and uninformed intensities. We note, however, that it is more accurate in determining the Poisson intensities than it is in



### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

determining the Bernoulli probabilities. Any inaccuracies are largely driven by a mis-estimation in the probability of news. Figure 3.9 shows the performance of CPIN (complete-linkage). We show that aside from a few outliers CPIN is generally good at estimating the Poisson intensities. Its estimation of the Bernoulli probabilities are less accurate than YZ-LK-PIN.

In figures 3.10 to 3.12 we analyze the effect of altering  $k$  on PIN parameter estimation. For each realization set  $(a, b, c)$  generated from three independent  $U[0, 1]$  random variables, we estimate PIN parameters using both  $k = 2,500$  and  $k = 5,000$ . Theoretically the estimated parameters in both scenarios should be exactly the same. Results show that whilst YZ-LK-PIN and CPIN (complete-linkage) perform satisfactorily, YZ-EHO-PIN results leave something to be desired.

Table 3.1 documents the mean squared error (MSE) and the mean absolute error (MAE) between theoretical PIN and estimated PIN for the two scenarios  $k = 2,500$  and  $k = 5,000$ . It can be seen that YZ-LK-PIN is the most accurate, but CPIN is not much worse. On the other hand YZ-EHO-PIN is considerably worse.

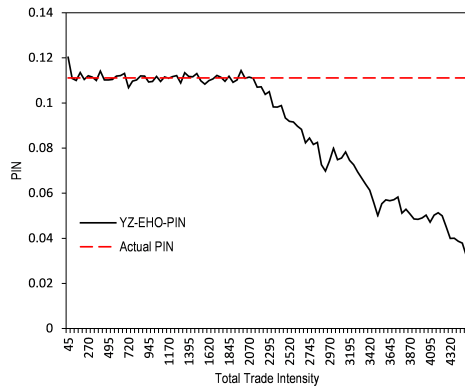
We conclude that YZ-LK-PIN is only marginally better in terms of accuracy. However, CPIN is significantly better in terms of speed.

### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

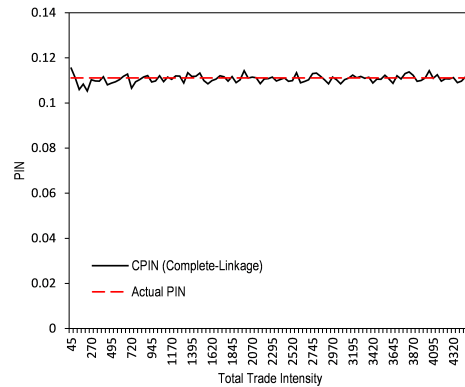
**Figure 3.4:** Test of Estimation Bias when increasing Trade Intensity

We test PIN estimation for different trade intensity scenarios. The theoretical PIN value remained the same at  $PIN = 0.1111$  ( $\alpha = \frac{1}{2}$ ,  $\delta = \frac{1}{2}$ ,  $\mu = 0.2k$ ,  $\varepsilon_B = 0.4k$ ,  $\varepsilon_S = 0.4k$ ); we vary  $k$  to increase trades. For each trade intensity scenario, we run YZ-EHO-PIN, YZ-LK-PIN and CPIN estimates 50 times. Below we plot the mean estimate for each scenario.

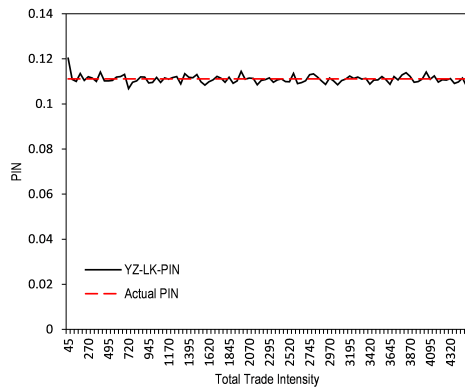
It is clear that YZ-EHO-PIN grossly underestimates the actual PIN when the total number of trades per day exceeds 2000. Both YZ-LK-PIN and CPIN (Complete-Linkage) perform well, and do not seem to have any bias with respect to trade intensity. Average-linkage and Ward's method produces some bias when total trade intensity is low.



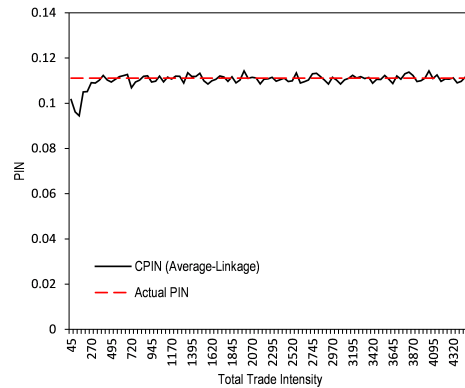
(a) YZ-EHO-PIN



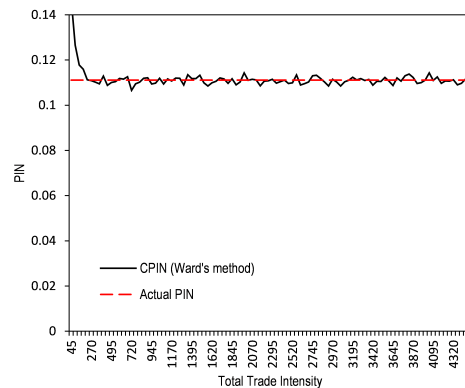
(b) CPIN (Complete-Linkage)



(c) YZ-LK-PIN



(d) CPIN (Average-Linkage)



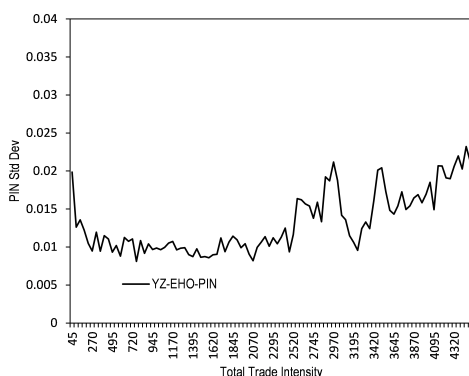
(e) CPIN (Ward's Method)

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

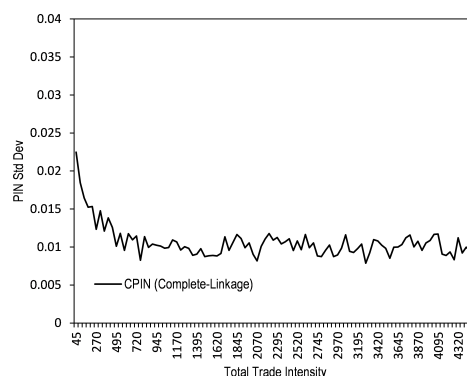
**Figure 3.5:** Test of Estimation Variance when increasing Trade Intensity

We test PIN estimation for different trade intensity scenarios. The theoretical PIN value remained the same at  $PIN = 0.1111$  ( $\alpha = \frac{1}{2}$ ,  $\delta = \frac{1}{2}$ ,  $\mu = 0.2k$ ,  $\varepsilon_B = 0.4k$ ,  $\varepsilon_S = 0.4k$ ); we vary  $k$  to increase trades. For each trade intensity scenario, we run YZ-EHO-PIN, YZ-LK-PIN and CPIN estimates 50 times. Below we plot the mean estimate for each scenario.

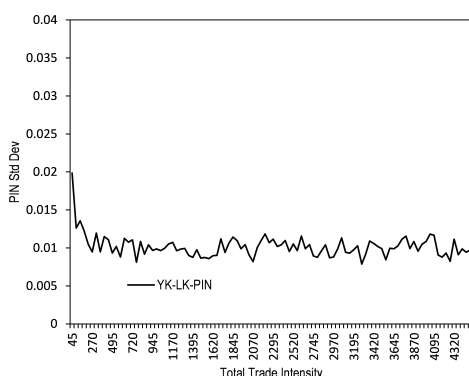
All methods show some variance in estimates when total trade intensity is low; this is most pronounced with using average-linkage clusters. CPIN Ward's method and complete-linkage are comparable to YZ-LK-PIN and superior to YZ-EHO-PIN.



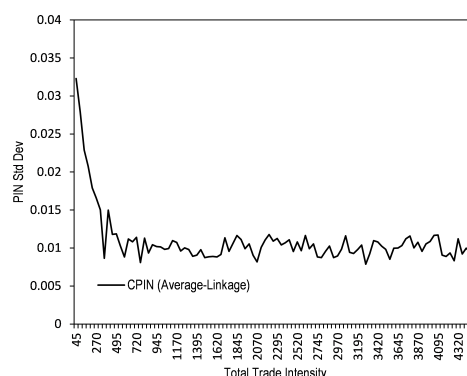
(a) YZ-EHO-PIN



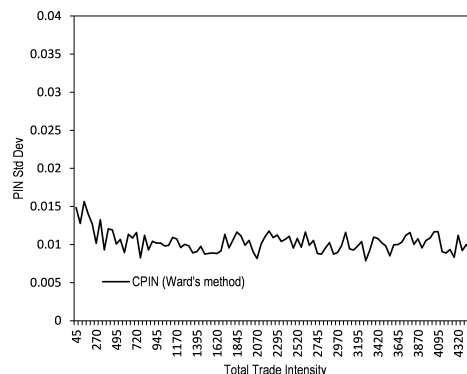
(b) CPIN (Complete-Linkage)



(c) YZ-LK-PIN



(d) CPIN (Average-Linkage)



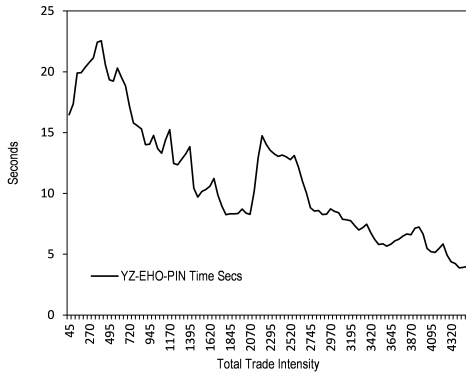
(e) CPIN (Ward's Method)

### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

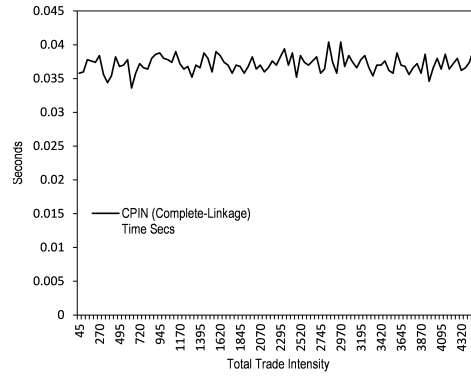
**Figure 3.6:** Test of Estimation Time when increasing Trade Intensity

We test PIN estimation for different trade intensity scenarios. The theoretical PIN value remained the same at  $PIN = 0.1111$  ( $\alpha = \frac{1}{2}$ ,  $\delta = \frac{1}{2}$ ,  $\mu = 0.2k$ ,  $\varepsilon_B = 0.4k$ ,  $\varepsilon_S = 0.4k$ ); we vary  $k$  to increase trades. For each trade intensity scenario, we run YZ-EHO-PIN, YZ-LK-PIN and CPIN estimates 50 times. Below we plot the average time it took for one estimate.

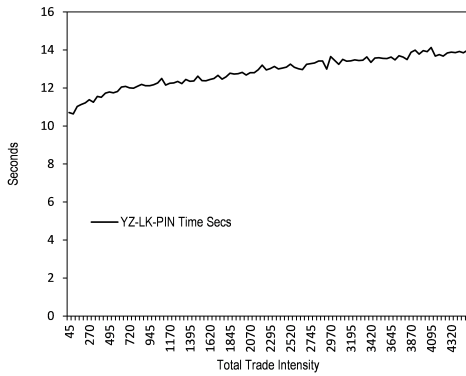
CPIN methods are much faster, taking on average 0.037 seconds to compute a PIN value. YZ-EHO-PIN took as long as 22 seconds to compute. The gradual reduction is due to failure to converge when number of trades are high. YZ-LK-PIN took an increasingly longer time to compute as the number of trades went up (from approx 11 seconds to 14 seconds). This was conducted on R 2.15.2 "Trick or Treat" version 64 bit using a Dell Optiplex 980 Intel i5 650 @ 3.19 GHz. YZ-EHO-PIN and YZ-LK-PIN were estimated using the `optim()` function. CPIN was estimated using the `hclust()` function.



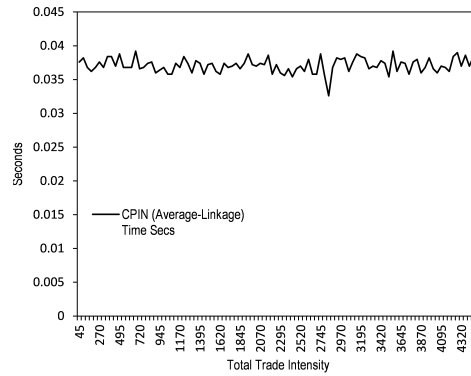
(a) YZ-EHO-PIN



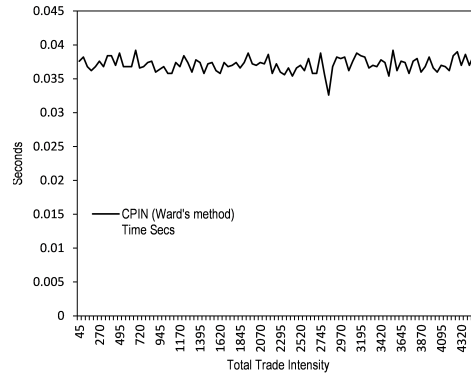
(b) CPIN (Complete-Linkage)



(c) YZ-LK-PIN



(d) CPIN (Average-Linkage)



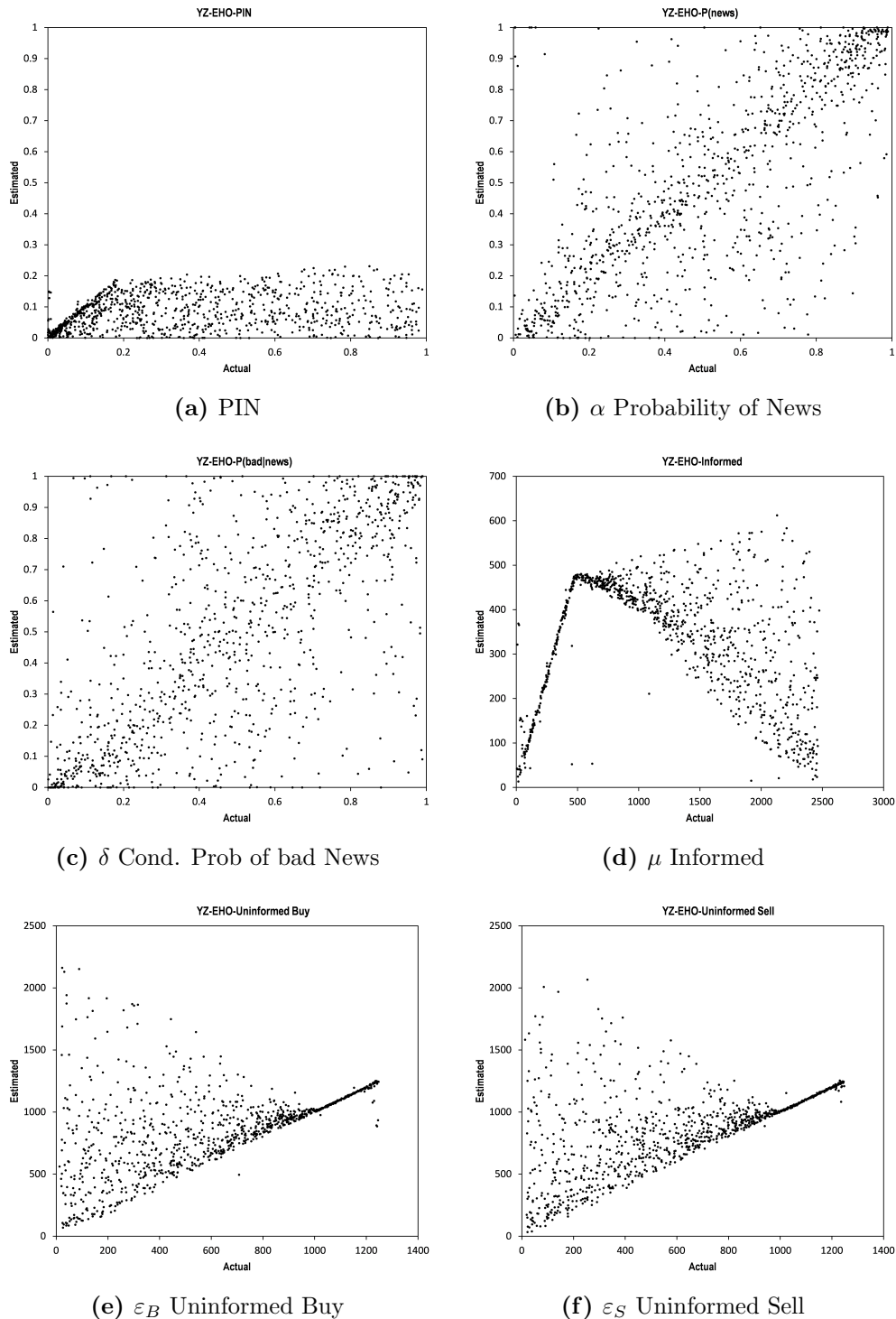
(e) CPIN (Ward's Method)

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

**Figure 3.7:** YZ-EHO-PIN Actual vs. Estimates

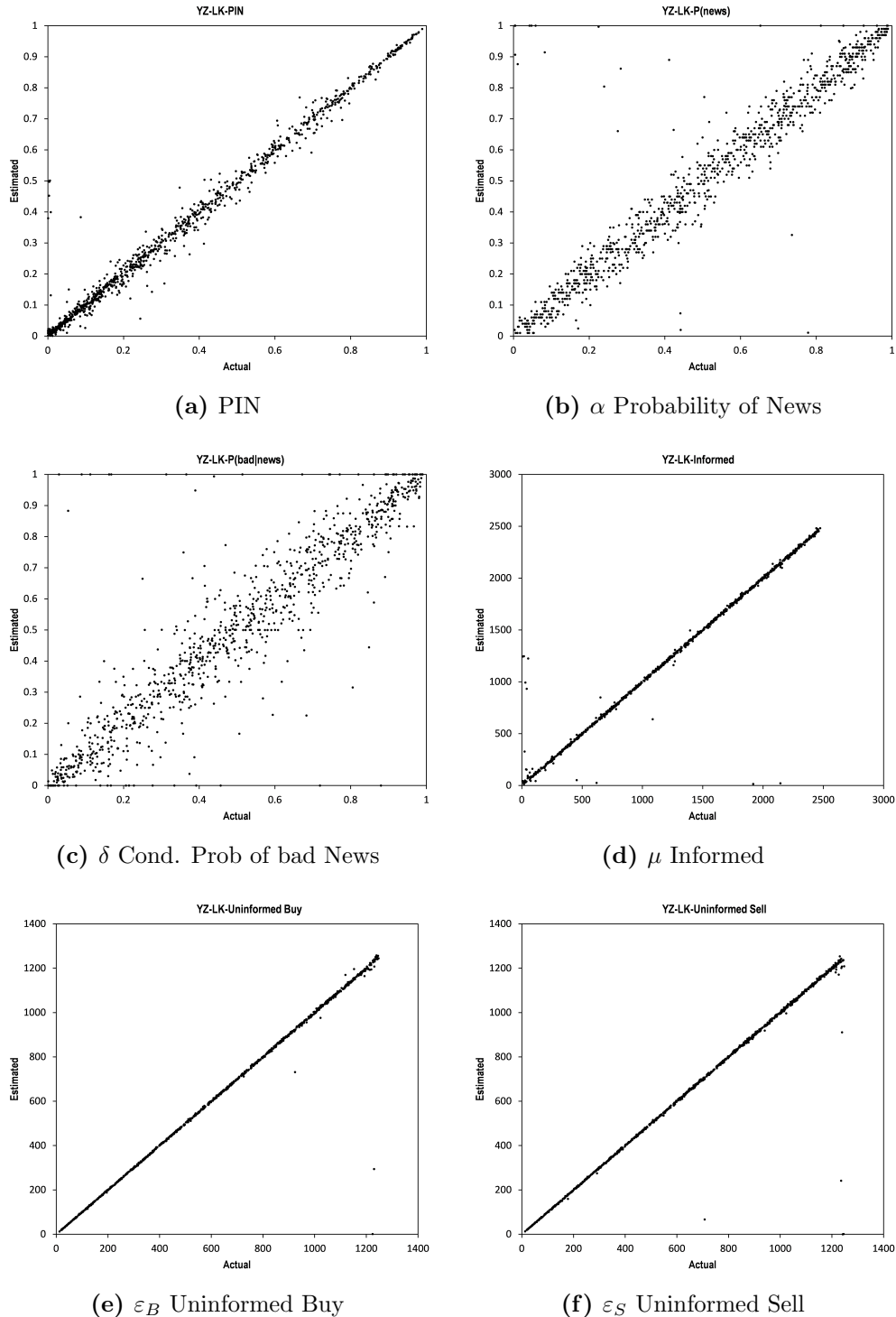
We test the accuracy of the YZ-EHO-PIN estimates by randomly generating a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O’Hara’s (2002) PIN model setup. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)/2k$ ,  $\varepsilon_S = (1 - c)/2k$ ; We let  $k = 2500$  and vary  $a, b, c$  bounded by 0 and 1.) This way, we can compare the actual (or theoretical) PIN value with the estimated PIN.



### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

**Figure 3.8:** YZ-LK-PIN Actual vs. Estimates

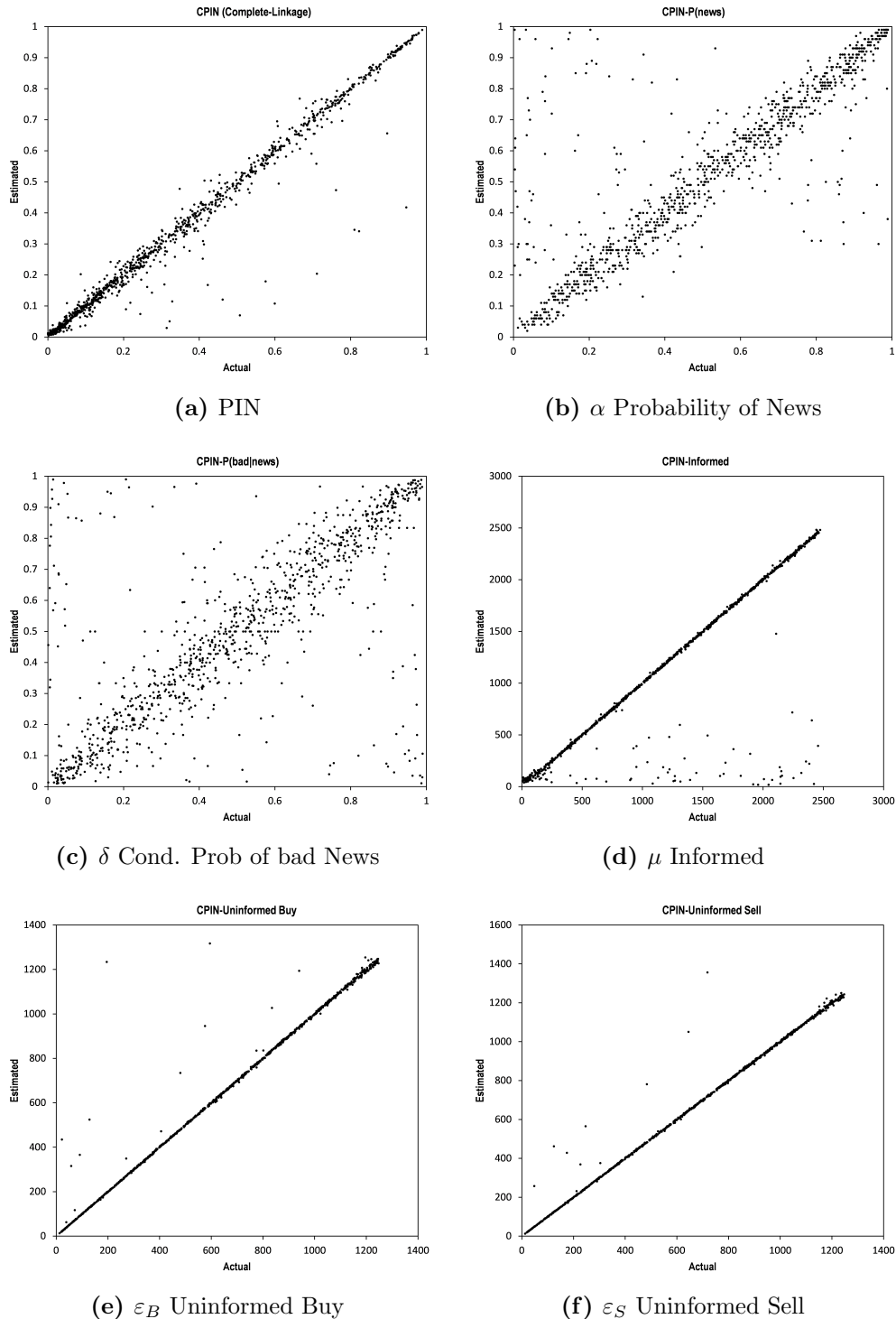
We test the accuracy of the YZ-LK-PIN estimates by randomly generating a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O'Hara's (2002) PIN model setup. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)/2k$ ,  $\varepsilon_S = (1 - c)/2k$ ; We let  $k = 2500$  and vary  $a, b, c$  bounded by 0 and 1.) This way, we can compare the actual (or theoretical) PIN value with the estimated PIN.



### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

**Figure 3.9:** CPIN Actual vs. Estimates

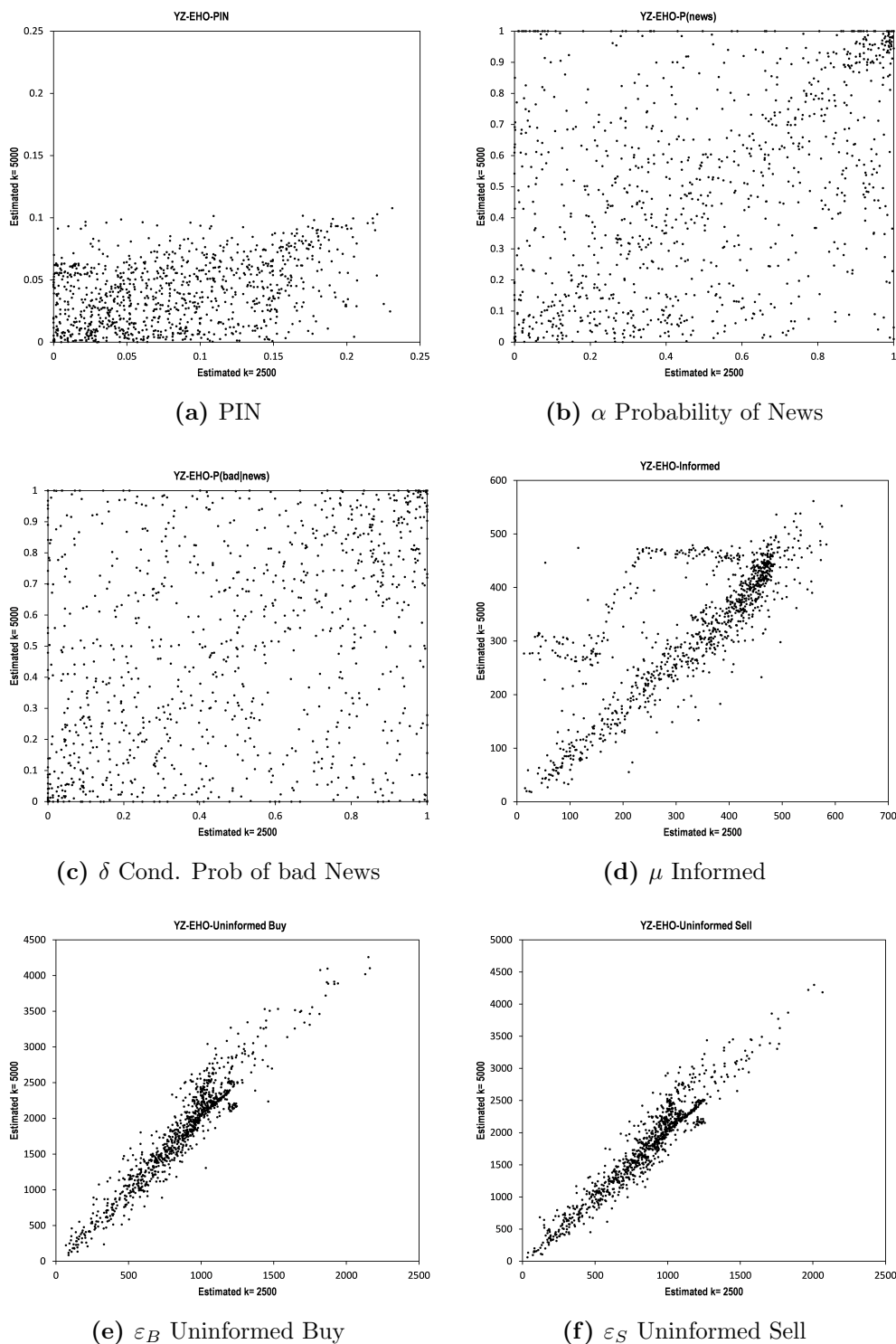
We test the accuracy of the CPIN (Complete-Linkage) estimates by randomly generating a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O'Hara's (2002) PIN model. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)/2k$ ,  $\varepsilon_S = (1 - c)/2k$ ; We let  $k = 2500$  and vary  $a, b, c$  bounded by 0 and 1.) This way, we can compare the actual (or theoretical) PIN value with the estimated PIN.



### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

**Figure 3.10:** YZ-EHO-PIN Estimates  $k = 2,500$  vs  $k = 5,000$

We show the impact of changing the total number of trades  $k$  on the estimated parameters. We randomly generate a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O'Hara's (2002) PIN model setup. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)k/2$ ,  $\varepsilon_S = (1 - c)k/2$ ) We let and vary  $a, b, c$  bounded by 0 and 1 and generate two sets of buy sell series, one where  $k = 2500$  and one where  $k = 5000$ , PIN parameter estimation should be the same for both cases. We plot scatter-plots to show if this is indeed the case.

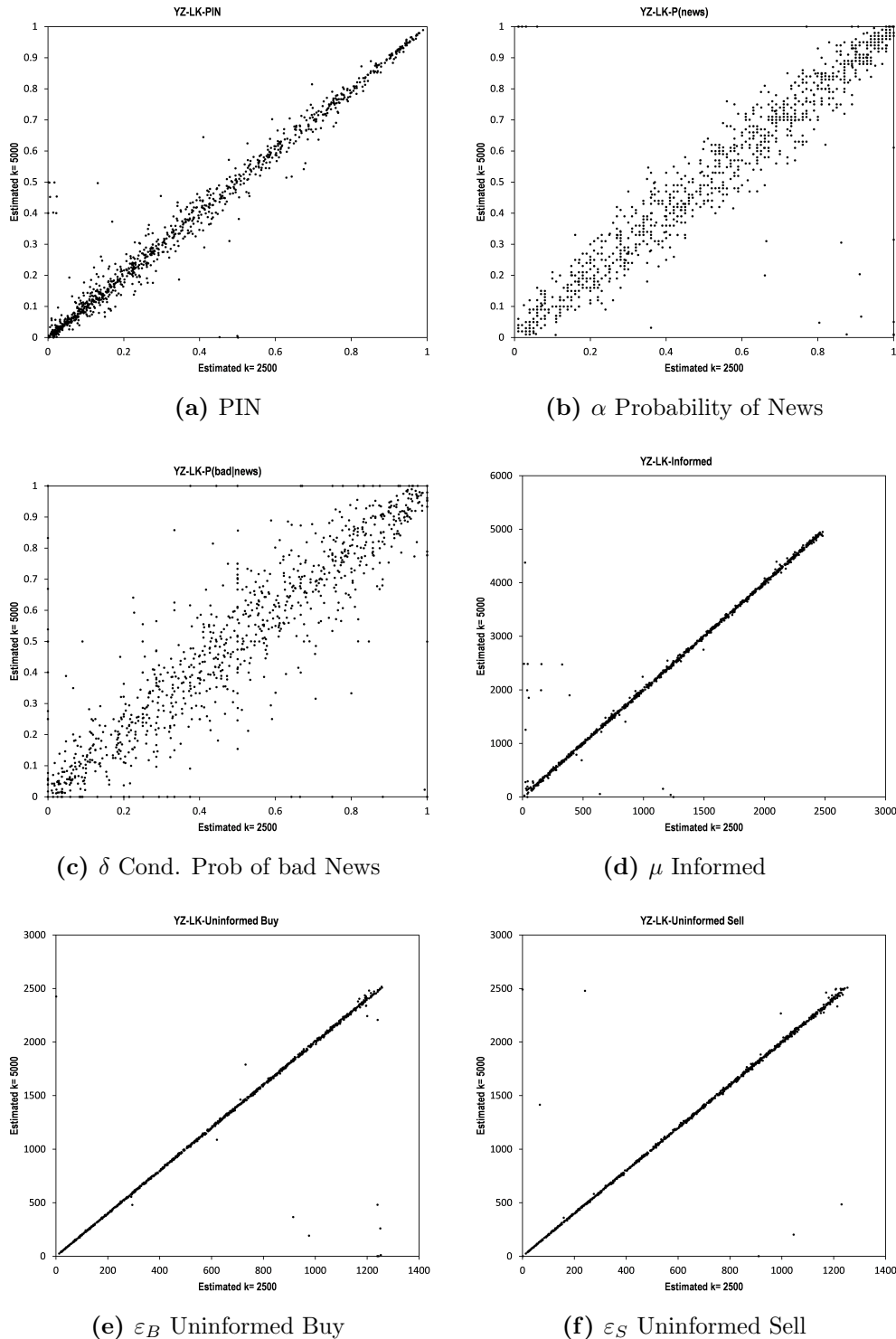




### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

**Figure 3.11:** YZ-LK-PIN Estimates  $k = 2,500$  vs  $k = 5,000$

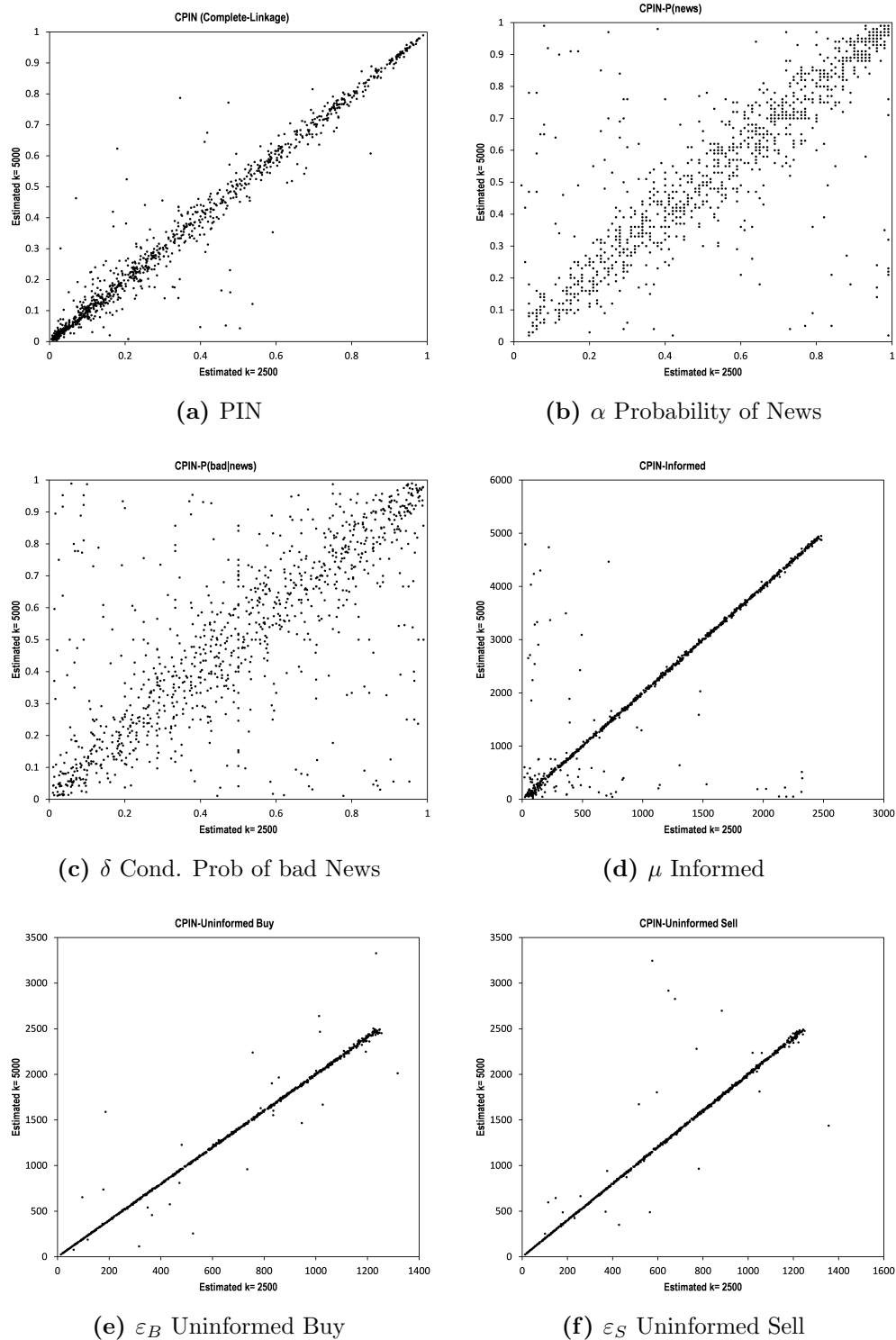
We show the impact of changing the total number of trades  $k$  on the estimated parameters. We randomly generate a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O'Hara's (2002) PIN model setup. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)k/2$ ,  $\varepsilon_S = (1 - c)k/2$ ) We let and vary  $a, b, c$  bounded by 0 and 1 and generate two sets of buy sell series, one where  $k = 2500$  and one where  $k = 5000$ , PIN parameter estimation should be the same for both cases. We plot scatter-plots to show if this is indeed the case.



### 3.4 A comparison between YZ-EHO-PIN, YZ-LK-PIN and CPIN variations

**Figure 3.12:** CPIN Estimates  $k = 2,500$  vs  $k = 5,000$

We show the impact of changing the total number of trades  $k$  on the estimated parameters. We randomly generate a set of 1000 different time-series of buys and sells governed by Easley, Hvidkjaer and O'Hara's (2002) PIN model setup. ( $\alpha = a$ ,  $\delta = b$ ,  $\mu = ck$ ,  $\varepsilon_B = (1 - c)k/2$ ,  $\varepsilon_S = (1 - c)k/2$ ) We let and vary  $a, b, c$  bounded by 0 and 1 and generate two sets of buy sell series, one where  $k = 2500$  and one where  $k = 5000$ , PIN parameter estimation should be the same for both cases. We plot scatter-plots to show if this is indeed the case.



### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

**Table 3.1:** PIN Estimation Error for  $k = 2,500$  and  $k = 5,000$

We simulate daily buy and sell time-series and test the accuracy of the estimated PIN with the theoretical PIN using mean squared error (MSE) and mean absolute error (MAE). We let  $\alpha$  - P(news),  $\delta$  - P(bad—news),  $\mu$  - informed intensity,  $\varepsilon_B$  - uninformed buy intensity and  $\varepsilon_S$  - uninformed sell intensity to be structured as follows:

$$\alpha = a; \quad \delta = b; \quad \mu = ck; \quad \varepsilon_B = \frac{1-c}{2}k; \quad \varepsilon_S = \frac{1-c}{2}k$$

$k$  resembles total trade intensity (we set it at 2,500 and 5,000). We randomize by giving setting a, b and c to be random variables  $\sim U[0, 1]$ .

Mean Squared Error						
k = 2500	PIN	P(news)	P(bad—news)	Informed	Uninformed	Uninformed
YZ-EHO	0.15201	0.04752	0.05814	1412692.9	174560.7	156800.0
YZ-LK	<b>0.00188</b>	<b>0.01280</b>	<b>0.01521</b>	<b>15284.0</b>	2427.2	4617.6
CPIN	0.00270	0.02471	0.04323	111842.3	<b>2409.4</b>	<b>1026.5</b>
Mean Absolute Error						
k = 2500	PIN	P(news)	P(bad—news)	Informed	Uninformed	Uninformed
YZ-EHO	0.27786	0.13989	0.16474	921.4	244.2	229.7
YZ-LK	<b>0.01709</b>	<b>0.04684</b>	<b>0.06713</b>	<b>19.1</b>	<b>4.8</b>	7.2
CPIN	0.01948	0.07359	0.10557	78.6	6.9	<b>5.1</b>
Mean Squared Error						
k = 5000	PIN	P(news)	P(bad—news)	Informed	Uninformed	Uninformed
YZ-EHO	0.18365	0.09989	0.10173	6987042.8	852071.1	793066.5
YZ-LK	0.00302	<b>0.01244</b>	<b>0.01533</b>	<b>72873.8</b>	36958.3	19938.8
CPIN	<b>0.00269</b>	0.02812	0.04736	452172.3	<b>13093.9</b>	<b>12861.3</b>
Mean Absolute Error						
k = 5000	PIN	P(news)	P(bad—news)	Informed	Uninformed	Uninformed
YZ-EHO	0.32411	0.22647	0.22443	2188.4	573.3	548.5
YZ-LK	<b>0.01894</b>	<b>0.04588</b>	<b>0.06799</b>	<b>42.1</b>	21.2	<b>12.8</b>
CPIN	0.01989	0.07886	0.11002	160.5	<b>13.4</b>	14.8

### 3.5 Employing CPIN as a starting value methodology for LK-PIN

It is clear from our simulations that Lin and Ke's (2011) method is extremely accurate in estimating PIN and resolves the estimation problems in Easley, Hvidkjaer and O'Hara (2002,2010). Our CPIN methodology is shown to be similar in terms of estimation accuracy but much faster as we bypass Yan and Zhang's (2006, 2012) initial value algorithm. However, it can be seen from the sum of squared residuals (SSR) table that Lin and Ke's method is slightly more accurate than ours. If speed is of concern, then we recommend users to simply use our CPIN (complete-linkage) methodology as it provides a reasonable level of accuracy (see figure 3.4, 3.9, 3.12 table 3.1). If accuracy is of concern, then we recommend running CPIN to determine the initial values of PIN parameters which can then be employed by Lin and Ke's MLE procedure. This method bypasses Yan and Zhang's (2006,2012) initial value exhaustion algorithm which is rather time-consuming.

Here we document the accuracy and speed of two methods: (1) YZ-LK-PIN - Lin and Ke's (2011) methodology with Yan and Zhang's (2006,2012) initial value algorithm and (2) C-LK-PIN - Lin and Ke's (2011) methodology with CPIN estimates as the initial values. Table 3.2 documents the errors and figure 3.12 plots the estimation time. It can be seen that C-LK is more accurate than YZ-LK in terms of estimating PIN and the uninformed intensities  $\varepsilon_B, \varepsilon_S$ . YZ-LK still maintains a marginally better accuracy in estimating the probability of bad news  $\delta$  and informed intensities  $\mu$ . However, since estimating PIN is of interest, C-LK is the better method. We find that our modified Lin and Ke method (C-LK) to be superior to the original Lin and Ke (2011) method in both speed and accuracy.

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

**Table 3.2:** Estimation Error between YZ-LK-PIN and C-LK-PIN

We simulate daily buy and sell time-series and test the accuracy of the estimated PIN with the theoretical PIN using mean squared error (MSE) and mean absolute error (MAE). We let  $\alpha$  - P(news),  $\delta$  - P(bad—news),  $\mu$  - informed intensity,  $\varepsilon_B$  - uninformed buy intensity and  $\varepsilon_S$  - uninformed sell intensity to be structured as follows:

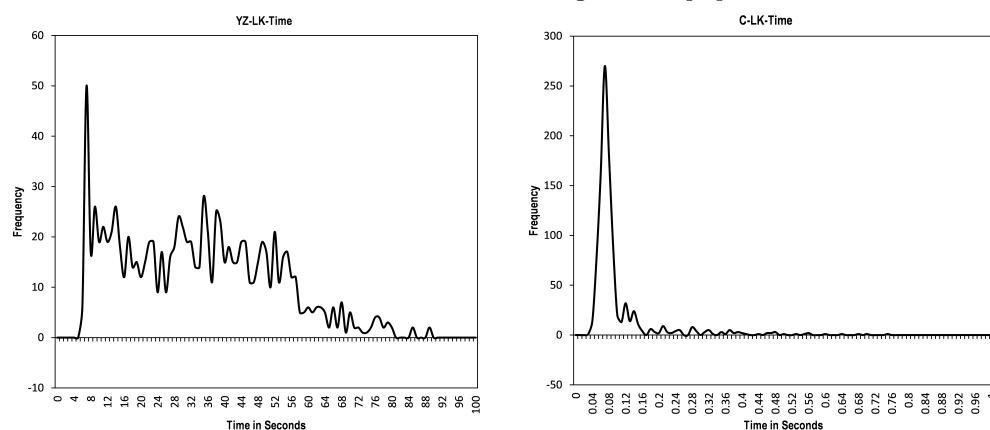
$$\alpha = a; \quad \delta = b; \quad \mu = ck; \quad \varepsilon_B = \frac{1-c}{2}k; \quad \varepsilon_S = \frac{1-c}{2}k$$

$k = 2,500$  resembles total trade intensity. We randomize by giving setting a, b and c to be random variables  $\sim U[0, 1]$ . It is clear than C-LK is marginally better than YZ-LK.

		Mean Squared Error				
		PIN	P(news)	P(bad   news)	Informed	Uninformed
YZ-LK		0.00164	0.01589	<b>0.01674</b>	<b>32699.3</b>	1578.8
C-LK		<b>0.00105</b>	<b>0.01437</b>	0.01748	34543.5	<b>99.9</b>
		Mean Absolute Error				
		PIN	P(news)	P(bad   news)	Informed	Uninformed
YZ-LK		0.01642	<b>0.04997</b>	<b>0.07105</b>	<b>27.2</b>	3.9
C-LK		<b>0.01528</b>	0.05023	0.07175	28.7	<b>3.4</b>

**Figure 3.13:** C-LK and YZ-LK Estimation Time Distribution

We compare the speeds of our method (C-LK) vs Lin and Ke’s (2011) original method (YZ-LK). The mean estimation time for C-LK is 0.096 seconds whilst the mean estimation time for YZ-LK is 32.74 seconds. This was conducted on R 2.15.2 ”Trick or Treat” version 64 bit using a Dell Optiplex 980 Intel i5 650 @ 3.19 GHz.



(a) YZ-LK estimation time

(b) C-LK estimation time

## 3.6 Classification of Good, Bad and No News Days

One clear advantage CPIN has over traditional PIN methods is its ability to explicitly classify good, bad and no news days. This cannot be achieved with YZ-EHO-PIN or YZ-LK-PIN methods. We can conduct ex-post analysis to test the accuracy of CPIN classification. Since CPIN is derived purely from daily buy and sell initiated trades, it is of interest to see if trading variables such as returns and volume are also significantly different between classifications. For example, one would expect CPIN classification of good (bad) news days should have significant positive (negative) returns.

We use tick history from 24 Korean stocks to conduct our empirical analysis. Trade initiation is classified using the Ellis, Michaely and O'Hara (2000) method<sup>1</sup>. Our dataset begins from January 2007 and ends at December 2012.

It is well known that electronic trading has proliferated over recent years; the number of trades in 2012 is significantly larger than the number of trades in 2007. To run CPIN over the whole sample set would be unwise as the order imbalance in recent years would be much larger than earlier years simply due to a growth in the number of trades. Therefore we only run CPIN for samples of 60 days and we run overlapping CPIN estimation across the full sample. This means firstly we run the CPIN procedure from day 1 to day 60; this will provide us with a  $60 \times 1$  classification vector of  $\{-1, 0, 1\}$  where -1 is a bad day, 0 is a no news day and 1 is a good news day. Then we run CPIN again for day 2 to day 61; ; this will provide us with another  $60 \times 1$  classification vector of  $\{-1, 0, 1\}$ . We continue doing this until we reach the end of the sample, i.e., day N-59 to day N. We discard the first and last 59 days. All the remaining days will have exactly 60 classifications (which are either -1, 0 or +1). The average of the 60 classification will give us a reading on that particular day; let us denote it  $z_t$ . If the average classification  $z_t$  is between  $(\frac{1}{3}, 1]$  then it is a good news day. If the average classification  $z_t$  is between  $[-1, -\frac{1}{3})$  then it is a bad news day. And if  $z_t$  is between  $(-\frac{1}{3}, \frac{1}{3})$  then it is a no news day.

---

<sup>1</sup>All trades executed at the ask quote are classified as a buy initiation. All trades executed at the bid quote are classified as a sell initiation. All other trades are categorized by the tick rule.

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

We perform two empirical tests in our study. Firstly, we conduct ex-post analysis on the classification to test differences in trading behavior between good news days, no news days and bad news days. Secondly, we want to see if the average classification time-series holds any sort of autoregressive features. The former may validate on how sensible our cluster method is in terms of clustering empirical data and the latter can test the speed of news or information dissemination in the market place.

From the average classification time-series we are able to define each day as being either good news, bad news or no news. From this we conduct a series of dummy variable regressions.

$$Vol_t^{Buy} = \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t$$

$$Vol_t^{Sell} = \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t$$

$$Spread_t = \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t$$

$$Volatility_t = \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t$$

$$Return_t = \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t$$

where  $Vol^{Buy}$  is daily buy initiated volume,  $Vol^{Sell}$  is daily sell initiated volume,  $Spread$  is the average daily relative spread,  $Volatility$  is the realized volatility derived from the sum of absolute quote revisions using the changes to the mid-price (to eliminate bid-ask bounce) and  $Return$  is the simple intra-day return. Table 3.3 shows a single stock example (POSCO) and table 3.4 provides aggregate summary across all 24 stocks.

From table 3.4, it is sensible and not surprising to find that buy initiated volume is significantly related to good news but not related to bad news. Likewise we find sell initiated volume is stronger in bad news days and not significant in good news days. Since our classifications were based on number of trades, it is useful to see that the volume data is also consistent with it. We find that the buy initiated volume to be stronger on a good news day than sell initiated

### 3.6 Classification of Good, Bad and No News Days

---

volume on a bad news days. This means more volume is being traded on good news days than on bad news days. We find this to be consistent with the short sale restriction effect. When there is good news, all the informed traders whether they hold the stock or not, will buy into the stock, driving up the volume. However, when there is bad news, only the informed traders that have the stock can sell, other informed traders are not allowed to short, therefore the sell initiated volume is lower on a bad news day than buy initiated volume is on a good news day.

From our ex-post analysis, we note that returns are also statistically significant with a positive (negative) sign on good (bad) news days. This makes sense and validates our CPIN classification method to be sensible. The bad (good) news group, classified using daily buy/sell initiations, indeed is reflected with bad (good) returns.

From table 3.4, we do not find any significant increase/decrease in spreads on good news days or bad news days. We do find realized intraday volatility to be higher in both good news days and bad news days when compared to no news days. This is also sensible as it suggests that intraday volatility is higher on news days than on no-news days. The positive relationship between news and volatility is well known. For example, Ross (1989) show price volatility to be positively correlated with information arrival. Our sensible ex-post results show that CPIN classification is reasonable.

Using the average classification time-series  $z_t$ , we test its autoregressive properties.  $z_t$  is a Real number between  $[-1, +1]$  inclusive.

$$z_t = \beta_0 + \sum_{i=1}^{10} \beta_i z_{t-i} + \varepsilon_t$$

Table 3.5 and 3.6 documents our results. In table 3.5, we run the autoregression on POSCO's average classification  $z_t$  from January 2007 to December 2012. Autoregressive lags 1 and 2 are significant at the 5% level. We run this for all stocks in our sample and tabulate aggregate results in table 3.6. Results show that 96% of all stocks in our sample exhibit statistical significant in the first two lags. This means that good (bad) news days are likely to result in further good (bad) news days. In a way, this indicates a certain amount of momentum in the daily



### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

information flow process.

**Table 3.3:** Trading Behavior in Good New and Bad News Days (POSCO 005490 KS)

We use CPIN to classify individual days into (1) good news days (2) bad news days and (3) no news days. We then regress daily volume, daily average relative spreads, daily realized volatility and daily returns on dummy variables indicating good news days and bad news days. We wish to test the differences in trading behavior between CPIN classified good news days versus bad news days. Regression equations are

$$\begin{aligned}
 Vol_t^{Buy} &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Vol_t^{Sell} &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Spread_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Volatility_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Return_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t
 \end{aligned}$$

where  $Vol_t^{Buy}$  and  $Vol_t^{Sell}$  are daily buy and sell initiated volume respectively.  $Dum$  refers to dummy variables.

Regress on Buy Volume						
	Coeff	T stat	p-value			Adj $R^2$
constant	6025.6	44.49	0.000	***		75.3%
Good	4119.3	15.61	0.000	***		
Bad	135.0	0.55	0.581			
Regress on Sell Volume						
	Coeff	T stat	p-value			Adj $R^2$
constant	6595.3	46.91	0.000	***		77.6%
Good	-607.0	-2.22	0.027	**		
Bad	4387.5	17.30	0.000	***		
Regress on Relative Spreads						
	Coeff	T stat	p-value			Adj $R^2$
constant	1.76E-03	20.91	0.000	***		31.2%
Good	4.06E-06	0.02	0.980			
Bad	-2.85E-04	-1.88	0.060	*		
Regress on Realized Volatility						
	Coeff	T stat	p-value			Adj $R^2$
constant	0.346	16.71	0.000	***		29.7%
Good	0.125	3.11	0.002	***		
Bad	0.113	3.02	0.003	***		
Regress on Returns						
	Coeff	T stat	p-value			Adj $R^2$
constant	6.51E-04	1.13	0.257			8.2%
Good	8.69E-03	7.76	0.000	***		
Bad	-6.78E-03	-6.54	0.000	***		

### 3.6 Classification of Good, Bad and No News Days

**Table 3.4:** Trading Behavior in Good New and Bad News Days (All stocks)

We use CPIN to classify individual days into (1) good news days (2) bad news days and (3) no news days. We then regress daily volume, daily average relative spreads, daily realized volatility and daily returns on dummy variables indicating good news days and bad news days. We wish to test the differences in trading behavior between CPIN classified good news days versus bad news days. Regression equations are

$$\begin{aligned}
 Vol_t^{Buy} &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Vol_t^{Sell} &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Spread_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Volatility_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t \\
 Return_t &= \beta_0 + \beta_G Dum_t^{Good} + \beta_B Dum_t^{Bad} + \varepsilon_t
 \end{aligned}$$

where  $Vol_t^{Buy}$  and  $Vol_t^{Sell}$  are daily buy and sell initiated volume respectively.  $Dum$  refers to dummy variables. Coeff refers to the average significant coefficient across 24 stocks. % Sign. refers to the percentage of stocks that had a p-value lower than 0.05 (i.e., significant at the 5% level).

Regress on Buy Volume	Coeff	% Sign.	Adj $R^2$
constant	<b>9,816.01</b>	<b>96%</b>	70.9%
Good	<b>7,744.54</b>	<b>100%</b>	
Bad	-632.68	54%	
Regress on Sell Volume	Coeff	% Sign.	Adj $R^2$
constant	<b>10,026.46</b>	<b>92%</b>	71.5%
Good	793.70	46%	
Bad	<b>5,338.60</b>	<b>100%</b>	
Regress on Relative Spreads	Coeff	% Sign.	Adj $R^2$
constant	<b>2.29E-03</b>	<b>100%</b>	49.3%
Good	-9.85E-05	25%	
Bad	-2.03E-04	50%	
Regress on Realized Volatility	Coeff	% Sign.	Adj $R^2$
constant	<b>0.451</b>	<b>100%</b>	45.4%
Good	<b>0.121</b>	<b>71%</b>	
Bad	<b>0.051</b>	<b>63%</b>	
Regress on Returns	Coeff	% Sign.	Adj $R^2$
constant	0.000	21%	7.9%
Good	<b>0.010</b>	<b>100%</b>	
Bad	<b>-0.008</b>	<b>100%</b>	

### 3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING

---

**Table 3.5:** Auto-regression on Classification (POSCO 005490 KS)

We run an auto-regression of up to lag 10 on the classification time-series and show that it is auto-correlated up to at least 2 days. This means the classification of good news days and bad news days is likely to be sticky.

	Coeff	T stat	p-value		Adj R2
constant	-0.014	-1.20	0.229		15.8%
<b>AR(1)</b>	<b>0.351</b>	<b>13.99</b>	<b>0.000</b>	<b>***</b>	
<b>AR(2)</b>	<b>0.053</b>	<b>1.99</b>	<b>0.047</b>	<b>**</b>	
AR(3)	0.021	0.80	0.424		
<b>AR(4)</b>	<b>0.049</b>	<b>1.86</b>	<b>0.063</b>	<b>*</b>	
AR(5)	-0.004	-0.13	0.893		
AR(6)	0.008	0.29	0.770		
AR(7)	0.010	0.37	0.710		
AR(8)	-0.005	-0.19	0.852		
AR(9)	0.034	1.28	0.201		
AR(10)	0.037	1.49	0.136		

**Table 3.6:** Auto-regression on Classification (All Stocks)

We run an auto-regression of up to lag 10 on the classification time-series and show that it is auto-correlated up to at least 2 days. This means the classification of good news days and bad news days is likely to be sticky.

	Coeff	% Sign.	Adj R2
constant	-0.0129	38%	17.2%
<b>AR(1)</b>	<b>0.2719</b>	<b>100%</b>	
<b>AR(2)</b>	<b>0.0979</b>	<b>96%</b>	
AR(3)	0.0225	38%	
AR(4)	0.0272	38%	
AR(5)	0.0193	25%	
AR(6)	0.0215	38%	
AR(7)	0.0126	21%	
AR(8)	0.0201	29%	
AR(9)	0.0066	13%	
AR(10)	0.0172	29%	

## **3.7 Conclusions**

In conclusion, we have compared several recent empirical methodologies for estimating the probability of informed trading, and have shown that Easley, Hvidkjaer and O'Hara's (2010) methodology to be inaccurate and Lin and Ke (2011) and Yan and Zhang's (2012) methods to be time-consuming. We have shown that our CPIN methodology provides both speed and accuracy to the user. We also illustrate the by-product of our methodology, which is the explicit classification of days. This ability provides researchers with the ability to identify good news, bad news and no news days, which was not possible with MLE approaches. Our methodology allows us to test the trading behavior of stocks around days with news and test autoregressive features of news.

### **3. A HIERARCHICAL AGGLOMERATIVE CLUSTERING APPROACH FOR ESTIMATING THE PROBABILITY OF INFORMED TRADING**

---

# References

- [1] Brown, S., Hillegeist, S.A. and K. Lo (2004) Conference calls and information asymmetry, *Journal of Accounting and Economics* 37, 343 - 366
- [2] Chae, J. (2005) Trading volume, information asymmetry, and timing information, *Journal of Finance* 60 413 - 442
- [3] Clark, P. (1973) A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* 41 135 - 155
- [4] Defays, D. (1977) An efficient algorithm for a complete link method, *The Computer Journal (British Computer Society)* 20, 364 - 366
- [5] Duarte, J., and L. Young (2009) Why is PIN priced? *Journal of Financial Economics* 91, 119 - 138
- [6] Easley, D., Hvidkjaer, S. and M. OHara, (2002) Is Information Risk a Determinant of Asset Returns?, *Journal of Finance* 10, 2185-2221
- [7] Easley, D., Hvidkjaer, S. and M. OHara, (2010) Factoring information into returns, *Journal of Financial and Quantitative Analysis* 45 - 2, 293 - 309
- [8] Easley, D., Kiefer, N.M. and M. O'Hara (1997) One day in the life of a very common stock, *The Review of Financial Studies* 10 - 3, 805 - 835
- [9] Easley, D., Kiefer, N., OHara, M., and J. Paperman, (1996) Liquidity, Information and Infrequently Traded Stocks, *Journal of Finance* 51, 1405-1436

## REFERENCES

---

- [10] Easley, D. and M.O'Hara (1992) Time and the process of security price adjustment, *Journal of Finance* 47, 577 -604
- [11] Lin, W.W. and W.C. Ke (2011) A computing bias in estimating the probability of informed trading, *Journal of Financial Markets* 14, 625 - 640
- [12] Ross, S.A. (1989) Information and volatility: the no-arbitrage martingale approach to timing and resolution irrelevancy, *Journal of Finance* 44, 1 - 17
- [13] Sokal, R. and C. Michener (1958) A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin* 38, 1409 - 1438
- [14] Ward, J.H. Jr. (1963) Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association* 58, 236 - 244
- [15] Yan, Y. and S. Zhang (2006) An improved estimation method and empirical properties of the probability of informed trading, working paper, University of Pennsylvania
- [16] Yan, Y. and S. Zhang (2012) An improved estimation method and empirical properties of the probability of informed trading, *Journal of Banking & Finance* 36, 454 - 467

## 4

# The Impact of Information Content and Illiquidity on Quote Revisions

1

**Abstract:** *In this chapter we study how information is impounded into prices through trades. It is widely believed that quote revisions are driven by the information content of trades (see Hasbrouck, 1991, Dufour and Engle, 2000, and Pascual, Escribano and Tapia, 2004). Dufour and Engle (2000) show that both trade duration and size of incoming order flow provide information on the level of informativeness of active trades. We extend upon this finding and suggest the possibility that illiquidity also plays a part in driving quote revisions. We hypothesize that quote revisions are not simply driven with incoming trades, but also driven by the existing liquidity of the order book as measured through spreads and depth. In order to test our hypothesis, we construct a sparse vector auto-regression (VAR) using tick data measurements (quote revisions, trades, durations, spreads and depths) to examine how trading attributes such as trade initiation, volume, durations, and order book attributes such as spreads and depths can impact subsequent quote revisions. For robustness, we employ adaptive lasso regularization (see Zou, 2006) which conducts VAR variable selection and parameter estimation in a single iteration. Our results show that aside from durations (Dufour and Engle, 2000) and trades (Hasbrouck, 1991), spreads*

---

<sup>1</sup>A variation of this chapter was presented at the Accounting and Finance Association of Australia and New Zealand (AFAANZ) in 1-3rd July, 2012, in Melbourne where it won the best finance paper award. It has also been presented at the Louis Bachelier Forum for Risk in Paris, France, 26th March 2013, the American Committee for Asian Economic Studies Financial Econometrics Group in Melbourne, 27th October 2012, and the 25th Australasian Finance and Banking Conference in Sydney, 18th December 2012



## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

*and depths also have a significant role in affecting price impact. Quote revisions are greater in periods when spreads are wider and depth is small. This result suggests quote revisions are not only driven by the information content of incoming trades but also by existing market illiquidity. As consequence to this, Dufour and Engle's VAR model is modified to allow for order book variables, and the price impact of trades is estimated using our new model.*

### 4.1 Introduction

An integral component in the field of market microstructure is the analysis of the information content of trades (see Pascual, Escribano and Tapia, 2004). It is generally accepted that market participants learn and update their beliefs and limit order quotes from incoming order flow. This process where passive traders adjust their positions from incoming active trades forms price discovery. In information asymmetry models (Kyle, 1985; Glosten and Milgrom, 1985; Easley and O'Hara, 1987; Admati and Pfleiderer, 1988; Foster and Viswanathan, 1993), the market is divided into two types of participants: informed and liquidity (uninformed) traders. Liquidity traders (and market makers) gradually revise quotes to reflect the private information from observing past trades by informed participants. In essence, price dynamics are determined through trade-by-learning mechanisms. For this reason, the private information is disseminated through the trading process described as a *tâtonnement* process, as prices are gradually adjusted to reflect the expectation of the true value of the security based on all current information available in the market place. In an environment with no additional information, prices converge to the true value in the long run via continuous trading and subsequent quote revisions by market makers. This adjustment in quotes is known as the price impact of a trade. Therefore, the study of price impact is a key component to understanding the dynamics of the price discovery process.

Hasbrouck (1991) is first to provide empirical evidence supporting the price impact of trades. He constructs a bi-variate VAR model using quote revisions (quote mid-point returns) and signed trades. He shows that quote revisions have a lagged response to trades and these lags can be attributed to inefficiencies inherent in the microstructure architecture, for example price dis-

creteness, inventory control effects and lagged information adjustment. Price impact can then be estimated as the sum of all quote revisions due to an order flow shock, i.e., the cumulative impulse response function of quote revisions from a unit trade shock. Dufour and Engle (2000) extend upon Hasbrouck (1991) by considering the role trade durations have to play in price impact. They find that in periods where the market is most active (shorter trade durations), price impact is higher, suggesting greater levels of information content and a higher percentage of informed trading. It is this research that confirms the saying '*fast trading is informed trading*'.

These studies were based on a reduced form approach to price impact, estimated through measuring the cumulative impulse response functions (CIRF) from VAR models that analyze the dynamics between price changes and trades. Pascual, Escribano and Tapia (2004) and Van Ness, Van Ness and Warr (2002) state that Hasbrouck's reduced form approach is superior to structural models based on Glosten and Harris (1988) and Huang and Stoll (1997). In de Jong, Nijman and Roëll (1996), they show that VAR models captured twice as much price impact to the structural Glosten (1994) model for the Paris Bourse. This is because structural models assume that price impact from a trade is instantaneous, whilst the VAR model accounts for possible lags.

However, we realize that Hasbrouck (1991) and Dufour and Engle (2000) consider only the incoming trades as being a factor for influencing the magnitude of quote revisions (and subsequently price impact). Their works consider trading attributes such as trade initiation, trade size, and trade duration. The implicit underlying assumption is that the cumulative quote revisions from the CIRF is reflective of the level of information content in incoming trades. We hypothesize that whilst the information content of incoming order flow does drive quote revisions, the existing status of the order book also drives quote revisions. Our study tests whether simple order book attributes, spreads and depth, are also able to influence quote revisions.

Furthermore, existing research by Hasbrouck (1991) and Dufour and Engle (2000) use limited data from the NYSE. Also both use only data from 1991. However, given advances in exchange technologies, we are interested in examining price impact in a modern Asian central limit order

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

book in 2012. The market examined in Hasbrouck (1991) and Dufour and Engle (2000) (i.e., NYSE) is a specialist quote-driven market, where designated specialists act as market makers and are required to trade to provide liquidity for the market. Therefore, there is a focus on the role of these specialists and how they revise quotes to reflect new information in these studies. In the modern day limit order book market, such as the Korea Exchange (KRX), specialists do not exist for large capitalization equities, instead the central limit order book itself is regarded as a competitive market maker. In essence, limit orders compete amongst each other in the central limit order book to make the market. Therefore, we probe to see if quote revisions are as reactive to trades in these circumstances as is suggested in information asymmetry models. Using high frequency timestamped trade-by-trade data (i.e. traded price, volume and respective bid and ask quotes) from the KRX, we construct our own VAR model for estimating price impact which extends upon Dufour and Engle (2000) by considering spreads and depth.

In specialist markets analyzed in Hasbrouck (1991) and Dufour and Engle (2000), the specialists extract private information from order flows and revise their quotes accordingly. A key question we propose to answer is whether the limit order book can extract information from order flows too?

We show that in a limit order book setting, changes in quote revisions reflect more than simply information content from trades. Market depths, spreads and durations may also play a part. This hasn't been analyzed previously.

Suppose two trades occur at time  $t_1$  and  $t_2$ . Let  $t_{1-}$  and  $t_{1+}$  be the time immediately before and after the trade at  $t_1$ . Likewise, let  $t_{2-}$  and  $t_{2+}$  be the time immediately before and after the trade at  $t_2$ . Therefore the chronological sequencing would be  $t_{1-} < t_1 < t_{1+} < t_{2-} < t_2 < t_{2+}$ .

Let  $m(t)$  be the log mid-point quote at time  $t$ . Then the quote revision after the first trade at  $t_1$  would be,

$$r_1 := m(t_{2-}) - m(t_{1-}) = [m(t_{1+}) - m(t_{1-})] + [m(t_{2-}) - m(t_{1+})]$$

We suggest that quote revisions can be driven by two components. The first component,  $m(t_{1+}) -$

$m(t_{1-})$ , is purely reflective on the size of the trade and the size of the market depth. A large market order may fill many limit orders on top of the book, resulting in a large change in the best bid or ask at time  $t_{1+}$ . This is based on the mechanics of the limit order book and is termed ‘walking down the order book’, and is not present in specialist markets.

The second component,  $m(t_{2-}) - m(t_{1+})$ , refers to the change in quotes between  $t_{1+}$  and  $t_{2-}$ . This occurs if there are new limit orders submitted during that period. If the trade at  $t_1$  contains significant information content, we would expect subsequent changes in quotes to materially impact the magnitude of this second component. It can be suggested that it is this component that most accurately reflects the information content of the trade. It is this component that is analyzed in Hasbrouck (1991) and Dufour and Engle (2000). However, we would still expect market depth and trade size to impact this component. Furthermore, if the bid ask spread is widened by a large trade, it is more likely that new orders will be submitted to undercut the spread. This affects the behavior of future quote revisions.

To the best of our knowledge, this is the first paper to use a regression shrinkage technique (adaptive lasso - least absolute shrinkage and selection operator) for a Hasbrouck-based endogenous trade & quote revision framework. Earlier works by Hasbrouck (1991) and Dufour and Engle (2000) did not put model specification and lag selection into consideration. Five lags are employed without discussion. In the VAR literature, it is common to use information criteria (i.e. Akaike IC, Schwarz Bayesian IC) for lag selection. However, these methods involving subset selection or stepwise techniques are computationally tedious and require a two step procedure, involving an exhaustive set of  $2^n$  combinations from  $n$  regressors. In this paper, we show how to employ adaptive lasso (Zou, 2006) in a VAR framework to conduct model selection and estimate simultaneously. Adaptive lasso is more robust in regards to avoiding spurious regressors, and we use it to verify our OLS estimators. Statistical significance in market microstructure has always been problematic given the large sample sizes applied in estimation. This is because as sample size increases, standard errors to OLS estimates are reduced, causing p-values to fall into significance. However, these factors may be spurious. It is well known that often OLS estimates provide great in-sample fit with many selected factors, however perform poorly out of

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

sample, i.e. low prediction accuracy due to the selection of irrelevant factors (Tibshirani, 1996). Furthermore, it is well known that adaptive lasso holds *oracle properties* whilst OLS does not (Zou, 2006). In simple terms, this means as the sample size increases the estimated subset of significant variables asymptotes to the true subset of variables for adaptive lasso. Therefore, we advocate the use of variable selection for microstructure research as it promotes sparsity, interpretability and lower prediction error.

Our results show that past trades have a significant positive contribution to quote revisions. Whilst the impact of trades decays as the lags lengthen, it is significant to at least 6 lags<sup>1</sup>. This proves that the characteristics documented on the NYSE in 1991 is still present in 2012 on the KRX. We also find that durations do not have a strong impact on price impact, instead we find that spreads and depth are much better at influencing price impact. Wider spreads and smaller depths have a positive impact on the price impact of a trade. This seems to suggest that whilst part of price impact may be driven by informed trades and information content, another component of price impact is driven by illiquidity in the order book. We conclude that price impact is a product of both demand and supply interactions. On the demand side: incoming trades and on the supply side: existing liquidity in the order book.

The remaining chapter is structured as follows. Section 4.2 provides a detailed summary on the literature review with regards to price impact. Section 4.3 briefly discusses the data we use (for details see chapter 2). Section 4.4 presents our price impact model, section 4.5 explains how adaptive lasso regularization works, section 4.6 explains the adaptive lasso estimation procedure and section 4.7 briefly discusses how we construct impulse responses. Section 4.8 provides our empirical results from model estimation and we conclude in section 4.9.

---

<sup>1</sup>After conducting adaptive lasso VAR, all 24 stocks showed significance up to 6 lags, and two-thirds showed significance up to 10 lags

## 4.2 Literature Review

In this section we discuss the existing literature on quote revisions and estimating the price impact of a trade.

Hasbrouck (1991) is first to provide empirical evidence supporting the price impact of trades. He constructs a bi-variate VAR model using *quote revisions* or quote mid-point returns and *signed volume* (and also signed trades). Using tick data and allowing for contemporaneous effects flowing from signed volume to quote revision, Hasbrouck (1991) shows that quote revisions have a lagged response to signed volume. This is concluded through statistically significant lags in the least squares bi-variate VAR model. These lags are attributed to inefficiencies inherent in the microstructure architecture, for example price discreteness, inventory control effects and lagged information adjustment. Hasbrouck (1991) documents that infrequently traded stocks have greater price impact than frequently traded stocks, and furthermore, that higher volume trades have greater price impacts, though this effect diminishes with size. This study and subsequent studies by Madhavan, Richardson and Roomans (1997) and Huang and Stoll (1997) find that trades (i.e. signed volume) are successful in explaining subsequent quote movements, suggesting that there is predictive power.

Hasbrouck's (1991) general specification for  $p$  lags is as follows,

$$r_t = \sum_{i=1}^p a_i r_{t-i} + \sum_{i=0}^p b_i x_{t-i} + \nu_{1,t}$$

$$x_t = \sum_{i=1}^p c_i r_{t-i} + \sum_{i=1}^p d_i x_{t-i} + \nu_{2,t}$$

where  $r_t$  denotes quote revisions and  $x_t$  denotes signed trades. Notice that trades can impact quote revisions contemporaneously but not vice versa.

Following from the idea that trades convey information, it is not unreasonable to hypothesize that other trade attributes, such as trade durations, are also informative and may play a role

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

in price formation. The role of trade durations was first suggested by Diamond and Verrecchia (1987). Drawing from assumptions by information asymmetry models, market participants are divided into informed and uninformed investors. Uninformed or liquidity trades are considered as uniform and random - i.e. decisions made exogenous to the market. On the other hand, informed investors are only willing to trade when they hold informative news on the stock. However, due to short-selling restrictions and constraints, informed investors are more able to buy on good news than short-sell on bad news, leading to greater trading intensity during good news periods and lower intensity during bad news periods. Therefore, Diamond and Verrecchia (1987) conclude durations to be negatively correlated with returns. Easley and O'Hara (1992) hold a similar view, but where longer durations represent no news, and that informed traders only enter into trades when news exist. In both papers, it is evident that duration conveys information.

However, traditional microstructure literature based on information asymmetry models commonly disregard time when examining the price discovery process. It was Dufour and Engle (2000) that first suggest the incorporation of durations to the Hasbrouck (1991) VAR model. In essence, they believe that durations are able to affect market price behavior.

Dufour and Engle's (2000) extension examines the impact of durations and isolates the deterministic effect of time, therefore  $b_i$  and  $d_i$  in Hasbrouck (1991) are re-parameterized as follows,

$$b_i = \gamma_i + \sum_{j=1}^q \lambda_{j,i}^r D_{j,t-i} + \theta_i \ln(T_{t-i})$$
$$d_i = \beta_i + \sum_{j=1}^q \lambda_{j,i}^x D_{j,t-i} + \delta_i \ln(T_{t-i})$$

where  $T_{t-i}$  denote trade durations between trade  $x_{t-i}$  and  $x_{t-i-1}$  and  $D_{j,t-i}$  are a set of dummy variables for pre-determined intraday intervals to capture any diurnal effects. However, they find most of the diurnal dummies are not significantly different from zero. Their results show only trades in the first 30 minutes of trading had a different impact to the remaining trades throughout the day. Moreover, the model is truncated at 5 lags and is estimated via *ordinary*

*least squares*. Consequently, the Dufour and Engle (2000) model is simplified to the following,

$$r_t = \sum_{i=1}^5 a_i r_{t-i} + \lambda_{open}^r D_t x_t + \sum_{i=0}^5 [\gamma_i^r + \delta_i^r \ln(T_{t-i})] x_{t-i} + \nu_{1,t}$$

$$x_t = \sum_{i=1}^5 c_i r_{t-i} + \lambda_{open}^x D_{t-1} x_{t-1} + \sum_{i=1}^5 [\gamma_i^x + \delta_i^x \ln(T_{t-i})] x_{t-i} + \nu_{2,t}$$

where  $r_t$  denotes quote revisions,  $x_t$  denotes signed trades,  $D_t$  denotes the dummy variable which is 1 in the first 30 minutes of trading and zero otherwise and  $T_t$  denotes trade durations in seconds. Their findings show that shorter durations are related to larger quote revisions and stronger positive trade autocorrelations. For example, when a buy order is executed immediately after a previous order, it is more likely that it will be followed by another buy order. This is suggestive that trades during periods of excessive trading activity tend to impact quote revisions greater than periods of relative inactivity where durations are larger. Similar to Hasbrouck (1991), they find significant positive autocorrelation between signed trades. Dufour and Engle (2000) provides VAR coefficients for Fannie Mae (FNM) over a 62 day trading period from November 1990 to January 1991 using the TORQ database. They find all 5 lags that they employ in their model to be statistically significant for past trades and 4 lags to be significant for past quote revisions in the trade equation ( $x_t$ ). However, only lag 1 and 5 are significant for durations. In the quote revision equation ( $r_t$ ), up to 5 lags are significant for past quote revisions and 3 lags for past trades. Lagged durations are significant up to lag 2 and the diurnal dummy for the open is significant. Chen, Li and Cai (2008) also find similar characteristics in the Chinese market, suggesting that similar characteristics exist between developed and emerging markets, i.e. between the New York Stock Exchange and the Shanghai Stock Exchange.

Interestingly, Grammig, Theissen and Wünche (2011) using data from the Xetra open limit orderbook at Frankfurt Stock Exchange (FSE-Xetra) discover conflicting evidence to Dufour and Engle (2000). They find that trade intensity is inversely related to trade informativeness. This is explained drawing upon the *crowding out effect* by Parlour (1998). It is suggested in low information asymmetry markets with decent market liquidity, the crowding out of limit orders



## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

by market orders by impatient participants would cause high trade intensity and lower trade durations. Therefore active markets are expected to have smaller price impacts. Their result, as opposed to Dufour and Engle (2000) seem to infer that fast markets may not necessarily have higher adverse selection risks, and fast trading is not necessarily informed trading. We test if this is true for our data sample, i.e., the Korea Exchange (KRX).

### 4.3 Data

The data we use has been discussed in detail in chapter 2. To reiterate, we use 24 large capitalization Korean stocks listed on the KRX with complete tick history from January 2007 to December 2012. Trade and best quote data is sourced from Thomson Reuters Tick History. For details on data processing and our sample of stocks, please refer to chapter 2.

The KRX is a continuous pure limit order market with price/time priority. During continuous trading, any buy or sell order entered at a price that is equal to the ask or bid in the central limit orderbook will execute immediately. Once trades are executed, the volume will be deleted from the central limit order book. If the order volume cannot be executed completely, due to its size, the remaining volume enters the queue as a limit order. In instances where a market order is traded against several existing limit orders, the exchange generates a trade record for each market order - limit order pair of executing orders. In those instances, we aggregate all multiple trade records generated by a single market order into a single trade record.

The Thomson Reuters database does not provide us with trade initiation data, therefore we use Ellis, Michaely and O'Hara's (2000) trade initiation algorithm. In this classification method all trades executed at the ask (bid) quote are classified as a buy (sell) initiation. All remaining trades are classified via the tick rule. Ellis, Michaely and O'Hara (2000) finds this approach superior to Lee and Ready's (1991) method.

## 4.4 The VARX model

Here we present our price impact model; we denote it VARX - a VAR with exogenous variables. Similar to Hasbrouck (1991) and Dufour and Engle (2000), in our model we consider the interactions between trades and quote revisions. However, we wish to determine the role of durations, spreads and depth in this process. If the order book illiquidity metrics (spreads and depth) can influence quote revisions, then we have shown that price impact is not solely driven by incoming order flow but how incoming order flow interactions with existing order book illiquidity. Conceptually, we regard the incoming order flow to be the demand side of the price formation whilst the state of the order book provides information on the supply side of the price formation.

Therefore to formalize, two endogenous variables are considered,  $r_t$  and  $\dot{v}_t$ . Quote revisions  $r_t$  are defined to be the movement or changes in the midpoint price determined through the best bid and ask price in the order book  $r_t = 100 \times (\log(q_{t+1}) - \log(q_t))$  where  $q_t = \frac{q_t^{bid} + q_t^{ask}}{2}$ . Instead of transaction price, the use of the midpoint price  $q_t$  eliminates the bid-ask bounce associated with using returns generated through traded prices. The scaling factor of 100 is consistent with Dufour and Engle (2000).

Trades  $\dot{v}_t = x_t * \log v_t$  is signed trade volume, i.e., log volume multiplied by trade initialization. The log transformation is applied to reduce the impact of extraordinarily large volumes in the dataset, this is suggested by Potters and Bouchaud (2003) and Hafner (2005) whilst studying statistical properties of the market impact of trades.

Several exogenous variables are considered. Firstly, we consider the impact of trade durations  $d_t$ , as it is shown in the literature (Diamond and Verrecchia, 1987, and Easley and O'Hara, 1992) that durations (trading intensity) have a negative (positive) effect price impact. The rationale is that trading intensity increases in periods of greater information, and therefore each trade would contain greater information content, and subsequently greater price impact. This was confirmed in Dufour and Engle (2000). Secondly, we consider the bid-ask spread right before the trade,  $s_t$ , which is a common measure of static liquidity. Thirdly, we consider the first level

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

depth of the order book right before the trade  $h_t|x_t$ . This measure of liquidity is conditional on trade direction  $x_t$ . For a buy trade, we consider the first level depth of the ask side, and for a sell we consider the first level depth of the bid side.

Pascual, Escribano and Tapia (2004) show a generalization of Hasbrouck (1991). Our model draws upon the same framework.

$$\begin{aligned} r_t &= \sum_{i=1}^{\infty} \gamma_i^{(1,1)} r_{t-i} + \sum_{i=0}^{\infty} \left[ \gamma_i^{(1,2)} + \beta_i^{(1,1)} d_{t-i} + \beta_i^{(1,2)} s_{t-i} + \beta_i^{(1,3)} h_{t-i}|x_{t-i} \right] \dot{v}_{t-i} + \varepsilon_t^{(1)} \\ \dot{v}_t &= \sum_{i=1}^{\infty} \gamma_i^{(2,1)} r_{t-i} + \sum_{i=1}^{\infty} \left[ \gamma_i^{(2,2)} + \beta_i^{(2,1)} d_{t-i} + \beta_i^{(2,2)} s_{t-i} + \beta_i^{(2,3)} h_{t-i}|x_{t-i} \right] \dot{v}_{t-i} + \varepsilon_t^{(2)} \end{aligned} \quad (4.1)$$

It can be interpreted that  $d_{t-i}$ ,  $s_{t-i}$  and  $h_{t-i}|x_{t-i}$  are control variables, that will impact the relationship between trades and quote revisions. This can be re-expressed in VARX format with  $p$  lags.

$$\begin{aligned} \begin{bmatrix} 1 & -\gamma_0^{(1,2)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} r_t \\ \dot{v}_t \end{bmatrix} &= \begin{bmatrix} \gamma_1^{(1,1)} & \gamma_1^{(1,2)} \\ \gamma_1^{(2,1)} & \gamma_1^{(2,2)} \end{bmatrix} \begin{bmatrix} r_{t-1} \\ \dot{v}_{t-1} \end{bmatrix} + \dots + \begin{bmatrix} \gamma_p^{(1,1)} & \gamma_p^{(1,2)} \\ \gamma_p^{(2,1)} & \gamma_p^{(2,2)} \end{bmatrix} \begin{bmatrix} r_{t-p} \\ \dot{v}_{t-p} \end{bmatrix} \\ &+ \begin{bmatrix} \beta_0^{(1,1)} \\ 0 \end{bmatrix} d_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,1)} \\ \beta_1^{(2,1)} \end{bmatrix} d_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,1)} \\ \beta_p^{(2,1)} \end{bmatrix} d_{t-p} \dot{v}_{t-p} \\ &+ \begin{bmatrix} \beta_0^{(1,2)} \\ 0 \end{bmatrix} s_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,2)} \\ \beta_1^{(2,2)} \end{bmatrix} s_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,2)} \\ \beta_p^{(2,2)} \end{bmatrix} s_{t-p} \dot{v}_{t-p} \\ &+ \begin{bmatrix} \beta_0^{(1,3)} \\ 0 \end{bmatrix} h_t|x_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,3)} \\ \beta_1^{(2,3)} \end{bmatrix} h_{t-1}|x_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,3)} \\ \beta_p^{(2,3)} \end{bmatrix} h_{t-p}|x_{t-p} \dot{v}_{t-p} \\ &+ \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \end{aligned} \quad (4.2)$$

Let  $\mathbf{Y}_t = (r_t, \dot{v}_t)^T$ . Our VAR model is re-expressed as,

$$\begin{aligned} \mathbf{Y}_t &= \mathbf{A}_0 \mathbf{Y}_t + \dots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{B}_0 d_t \dot{v}_t + \dots + \mathbf{B}_p d_{t-p} \dot{v}_{t-p} \\ &+ \mathbf{C}_0 s_t \dot{v}_t + \dots + \mathbf{C}_p s_{t-p} \dot{v}_{t-p} + \mathbf{D}_0 (h_t|x_t) \dot{v}_t + \dots + \mathbf{D}_p (h_{t-p}|x_{t-p}) \dot{v}_{t-p} + \boldsymbol{\varepsilon}_t \end{aligned}$$

where,

$$\mathbf{A}_0 = \begin{bmatrix} 0 & \gamma_0^{(1,2)} \\ 0 & 0 \end{bmatrix} \quad \mathbf{B}_0 = \begin{bmatrix} \beta_0^{(1,1)} \\ 0 \end{bmatrix} \quad \mathbf{C}_0 = \begin{bmatrix} \beta_0^{(1,2)} \\ 0 \end{bmatrix} \quad \mathbf{D}_0 = \begin{bmatrix} \beta_0^{(1,3)} \\ 0 \end{bmatrix}$$

and for  $k = 1 \dots p$ ,

$$\mathbf{A}_k = \begin{bmatrix} \gamma_k^{(1,1)} & \gamma_k^{(1,2)} \\ \gamma_k^{(2,1)} & \gamma_k^{(2,2)} \end{bmatrix} \quad \mathbf{B}_k = \begin{bmatrix} \beta_k^{(1,1)} \\ \beta_k^{(2,1)} \end{bmatrix} \quad \mathbf{C}_k = \begin{bmatrix} \beta_k^{(1,2)} \\ \beta_k^{(2,2)} \end{bmatrix} \quad \mathbf{D}_k = \begin{bmatrix} \beta_k^{(1,3)} \\ \beta_k^{(2,3)} \end{bmatrix}$$

and  $\varepsilon_t$  is a white noise process with the covariance matrix  $E(\varepsilon_t \varepsilon_t^T) = \Sigma_\varepsilon$ .

Here we show the regression formulation for VAR models using the Kronecker tensor product. Let us define the following notations, where  $n$  is the number of total trade observations. VAR transformation is necessary allowing us to conduct adaptive lasso shrinkage.

$$\begin{aligned} \mathbf{Y}^* &= (\mathbf{Y}_{p+1}, \mathbf{Y}_{p+2}, \dots, \mathbf{Y}_n), \text{ and using the stack operator, } \mathbf{Y} = \text{vec}(\mathbf{Y}^*) \\ \mathbf{X}_t &= (\mathbf{Y}_t^T, \dots, \mathbf{Y}_{t-p}^T, d_t \dot{v}_t, \dots, d_{t-p} \dot{v}_{t-p}, s_t \dot{v}_t, \dots, s_{t-p} \dot{v}_{t-p}, (h_t | x_t) \dot{v}_t, \dots, (h_{t-p} | x_{t-p}) \dot{v}_{t-p})^T, \mathbf{X}^* = \\ &(\mathbf{X}_{p+1}, \dots, \mathbf{X}_n) \\ \mathbf{B} &= (\mathbf{A}_0, \dots, \mathbf{A}_p, \mathbf{B}_0, \dots, \mathbf{B}_p, \mathbf{C}_0, \dots, \mathbf{C}_p, \mathbf{D}_0, \dots, \mathbf{D}_p), \quad \beta = \text{vec}(\mathbf{B}) \\ \mathbf{U}^* &= (\varepsilon_{p+1}, \dots, \varepsilon_n), \quad U = \text{vec}(\mathbf{U}^*) \end{aligned}$$

Our model can be rewritten in concise matrix notation  $\mathbf{Y}^* = \mathbf{B}\mathbf{X}^* + U^*$  or as,

$$\mathbf{Y} = ((\mathbf{X}^*)^T \otimes I_2) \beta + U \equiv \mathbf{X} \beta + U, \quad U \sim (0, \Sigma_U = I_{n-p} \otimes \Sigma_\varepsilon) \quad (4.3)$$

The least squares estimator of  $\beta$  under the regression set up is,

$$\begin{aligned} \hat{\beta} &= ((\mathbf{X}^* (\mathbf{X}^*)^T)^{-1} \mathbf{X}^* \otimes I_2) \mathbf{Y} \\ \hat{\Sigma}_\varepsilon &= \frac{1}{n-k} (\mathbf{Y}^* - \hat{\mathbf{B}} \mathbf{X}^*) (\mathbf{Y}^* - \hat{\mathbf{B}} \mathbf{X}^*)^T \end{aligned}$$

$k$  is defined to be the number of beta coefficients. This is equivalent to the maximum likelihood estimator of VAR. As shown in Fuller (1996) and Lütkepohl (2005), VAR estimates are asymptotically consistent and normal<sup>1</sup>, and this results in our ability to empirically determine standard errors for VAR in model.

$$\widehat{Avar}(\hat{\beta}) = \left( \frac{1}{n} \Gamma^{-1} \otimes \hat{\Sigma}_\varepsilon \right)$$

<sup>1</sup>Given that the vector autoregressive processes are stationary processes for all  $t$  and that  $\varepsilon_t$  are independent with mean zero and covariance  $\Sigma_\varepsilon > 0$ , it can be shown that  $\lim_{n \rightarrow \infty} \mathbf{X}^* (\mathbf{X}^*)^T / n \rightarrow \Gamma$  where  $\Gamma$  is nonsingular or invertible (see Fuller, 1996 and Lütkepohl, 2005) and  $\hat{\beta}$  is asymptotically consistent  $p \lim_{n \rightarrow \infty} \hat{\beta} \rightarrow \beta$  and normal  $\sqrt{n}(\hat{\beta} - \beta) \rightarrow_d N(0, \Gamma^{-1} \otimes \Sigma_\varepsilon)$ .

## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

Section 4.5 to section 4.7 provide technical details on our model, in particular the implementation of regularization and estimating impulse response functions. Readers are welcome to skip these sections to the results in section 4.8 if they wish.

### 4.5 Regularization via Adaptive Lasso

We estimate our VARX model using adaptive lasso. It is a robust estimation method that performs variable selection and parameter estimation in a single step.

In the generalized multiple regression setting and also in our particular study using VAR, there is a two-fold objective (1) investigation of the relationship between the response and predictor variables and (2) prediction of future responses. As pointed out in Van der Kooij (2007), *ordinary least squares* fails in both respects, performing poorly in both model complexity and prediction accuracy. The main drawback with estimating VAR models is the number of variables and lags, leading to cases of over parametrization. This results in multicollinearity between different lagged variables as well as poor out of sample forecasts. In this paper, we use  $L_1$  shrinkage techniques. Selection via lasso (least absolute shrinkage and selection operator) is ideal because it selects the model and estimates parameters simultaneously. Compared to information criteria and stepwise based techniques, it is less computationally intense and more stable (Savin, 2010).

**Definition 1.** *The  $q$ -norm of a  $p \times 1$  vector  $\beta$  is denoted by  $\|\beta\|_q = (\sum_{i=1}^p |\beta_i|^q)^{\frac{1}{q}}$*

The loss function of ordinary least squares employs the squared Euclidean norm,

$$L^{LS}(\beta) = \|y - X\beta\|_2^2$$

The loss functions of lasso is simply a constrained version of least squares.

**Definition 2.** *The Lasso (Tibshirani, 1996) regression coefficients where the regression set up  $\hat{\beta} = \arg \min_{\beta} \|y - X\beta\|_2^2$  is subject to  $\|\beta\|_1 \leq t$  (constrained regression)*

This can also be expressed as an  $L_1$  regularized optimization problem,

$$\min_{\beta} L(\beta) = \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where  $\lambda \geq 0$  (penalized regression)

In essence, the Lasso shrinks the coefficients towards 0 as  $\lambda$  increases. Shrinkage is known to improve prediction accuracy due to the bias-variance trade-off. However, Fan and Li (2001) show unsatisfactory asymptotic features with Tibshirani (1996)'s Lasso estimator. Lasso was shown to be inconsistent in subset selection and biased in parameter estimation. Zou (2006) suggested adaptive Lasso as an improvement on the original framework. By allocating higher penalty for zero coefficients and lower penalty for nonzero coefficients, adaptive Lasso reduces estimation bias. Adaptive Lasso is known for consistency in variable selection and nonzero estimators are asymptotically normal.

**Definition 3.** *Adaptive Lasso (Zou, 2006) is a weighted  $L_1$  penalization method, imposing different shrinkage values for different parameters. We minimize the loss function below,*

$$\min_{\beta} L(\beta) = \|y - X\beta\|_2^2 + \lambda_n \sum_{j=1}^{param} \hat{\omega}_j \|\beta_j\|_1$$

where  $\lambda_n$  is the tuning parameter and  $\hat{\omega}_j = \|\hat{\beta}_j\|^{-\gamma}$  are the adaptive weights which are different for different parameters. We note  $\gamma \geq 0$ , where  $\gamma = 0$  is simply Tibshirani (1996)'s Lasso estimator.

Due to the fact that the loss function is convex, there are no issues with the ability to obtain a global minima. Zou (2006) and Ren and Zhang (2010) prove the *oracle property* for adaptive Lasso estimators.

**Theorem 1.** *The oracle property (Ren and Zhang, 2010) shows that adaptive Lasso VAR estimators have selection consistency and asymptotic normality.*

Let  $A = \{j : \beta_j \neq 0\}$  be the subset of interest. Selection consistency suggests convergence in probability of the estimated subset to the true subset.

## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

$$\lim_{n \rightarrow \infty} P(\hat{A}_n = A) = 1$$

Asymptotic normality exists for nonzero estimates,

$$\sqrt{n}(\hat{\beta}_{\hat{A}_n} - \beta_A) \rightarrow_d N(0, (\Gamma^{-1} \otimes \Sigma_\varepsilon)_A)$$

For proof, see Zou (2006) and Ren and Zhang (2010).

### 4.6 Adaptive Lasso Estimation

Ren and Zhang (2010) show how least angle regression (LARS) can be effectively utilized to compute adaptive lasso estimates, and their entire coefficient path. Its computational intensity is of the order  $O(np^2)$ , which is the same as a single *least squares* fit.

**Algorithm 1.** *LARS algorithm for adaptive lasso VAR (Ren and Zhang, 2010)*

1. Define  $x_j^* = x_j \widehat{\omega}_j$ ,  $j = 1, \dots, pk^2 + k$  (with inclusion of constant coefficients) where  $\widehat{\omega}_j = \|\widehat{\beta}^{ols}\|^{-\gamma}$
2. The LARS algorithm is used to compute the entire solution path of the Lasso.

$$\widehat{\beta}^{*(n)} = \arg \min_{\beta} \{\|y - X^* \beta\|_2^2 + \lambda_n \|\beta\|_1\}$$

3. Output  $\widehat{\beta}_j^{(n)} = \widehat{\beta}_j^{*(n)} \widehat{\omega}_j$ ,  $j = 1, \dots, pk^2 + k$

As is discussed in Zou (2006), tuning is an integral component in computation. Ren and Zhang (2010) suggest for each given  $\gamma$ , we search for the optimal  $\lambda_n$  using BIC criterion. Then grid search is used to find the optimal  $\gamma$ . Zou (2006) suggest using a 2-dimensional cross-validation to tune the adaptive lasso parameter pair  $(\gamma, \lambda_n)$ . Similar to Ren and Zhang (2010), Zou (2006) suggest for every given  $\gamma$  (in their paper  $\gamma \in \{0.5, 1, 2\}$ ) to use cross-validation along with the LARS algorithm to search for the optimal  $\lambda_n$ .

We write an adaptive lasso function faithful to Zou (2006) in R utilizing the R package

'LARS' developed by Hastie and Efron (2011) for computation and cycle through  $\gamma$  between 0.5 and 2 and use the BIC criterion for optimal  $\lambda_n$ .

**Extension: Standard Errors**

The computation of standard errors of non-zero lasso estimates is a topic for discussion. Tibshirani (1996) explains that it is difficult to obtain accurate estimates of standard errors given the non-linear and non-differentiable nature of the lasso estimates. In practice, it is popular to use lasso for variable selection to determine the best subset, and then reverting to least squares standard errors for that subset. Fan and Li (2001) show that local quadratic approximation (LQA) could be used to provide a sandwich formula for computing the covariance matrix of nonzero components of penalized estimates. In Zou (2006), it is briefly mentioned how to employ the sandwich formula to adaptive lasso standard errors.

Let  $A_n^*$  be the subset of nonzero penalized estimates, and let this be  $d$  variables. Let  $X_d^*$  be the regressor matrix which contains only the factors that have non-zero beta estimates. Let  $\Sigma(\beta) = \text{diag}(\frac{\hat{\omega}_1}{|\beta_1|}, \dots, \frac{\hat{\omega}_d}{|\beta_d|})$  Following Zou (2006)'s method, the estimated covariance matrix for adaptive Lasso  $\hat{\beta}^{(n)}$  in *Model 1* can be determined as follows, For the quote revision equation,

$$\widehat{cov}(\hat{\beta}_{A_n^*}^{(n)}) = \widehat{\Sigma}_{\varepsilon 1,1}(X_{A_n^*}^*(X_{A_n^*}^*)^T + \lambda_n \Sigma(\hat{\beta}_{A_n^*}^{(n)}))^{-1} X_{A_n^*}^*(X_{A_n^*}^*)^T (X_{A_n^*}^*(X_{A_n^*}^*)^T + \lambda_n \Sigma(\hat{\beta}_{A_n^*}^{(n)}))^{-1}$$

For the trade equation,

$$\widehat{cov}(\hat{\beta}_{A_n^*}^{(n)}) = \widehat{\Sigma}_{\varepsilon 2,2}(X_{A_n^*}^*(X_{A_n^*}^*)^T + \lambda_n \Sigma(\hat{\beta}_{A_n^*}^{(n)}))^{-1} X_{A_n^*}^*(X_{A_n^*}^*)^T (X_{A_n^*}^*(X_{A_n^*}^*)^T + \lambda_n \Sigma(\hat{\beta}_{A_n^*}^{(n)}))^{-1}$$

We use this approach to generate our standard errors.

## 4.7 Impulse Response Functions

In this section we examine some methods that could be applied for the generation of impulse response, which as is documented in Hasbrouck (1991) provides key insight on the magnitude and direction of market impact.



#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

For a reduced VAR(p) form model,

$$y_t = \sum_{i=1}^p A_i y_{t-i} + u_t$$

where  $A_1 = I_{2k+4} + \alpha\beta^T + \Gamma_1$ ,  $A_i = \Gamma_i - \Gamma_{i-1}$  and  $A_p = -\Gamma_{p-1}$ . Which can be reexpressed as VAR companion form,

$$Y_t = AY_{t-1} + U_t$$

where  $Y_t := [y_t, y_{t-1}, \dots, y_{t-p+1}]^T$  and  $U_t := [u_t, 0, \dots, 0]^T$  and,

$$A := \begin{bmatrix} A_1 & \dots & A_{p-1} & A_p \\ I_{2k+4} & & 0 & 0 \\ & \ddots & \vdots & \vdots \\ 0 & \dots & I_{2k+4} & 0 \end{bmatrix}.$$

Furthermore, it is noted that a VAR(1)  $\rightarrow$  VMA( $\infty$ ), via consecutive substitution.

$$Y_t = \sum_{i=1}^{t-1} A^i U_{t-1}$$

The impulse response function as a function of time and the innovation vector can be written as,

$$f(t; \delta) = JA^t J^T \delta$$

where  $J := [I_{2k+4}; 0; \dots; 0]$ , as only the first submatrix of  $A$  is useful.

It is important to derive confidence intervals for impulse responses to determine whether they are statistically meaningful. According to Lütkepohl and Reimers (1992), since VAR estimates are asymptotically normal  $\sqrt{T}vec[(\hat{A}_1, \dots, \hat{A}_p) - (A_1, \dots, A_p)] \rightarrow_d N(0, \Sigma)$ , then impulse responses  $\hat{\Phi}_n = \sum_{j=1}^n \hat{\Phi}_{n-j} \hat{A}_j \quad \forall n = 1, 2, \dots$  and accumulated impulse responses  $\hat{\Psi}_m = I_{2k+4} + \sum_{n=1}^m \hat{\Phi}_n$ , also have asymptotical normal distributions  $\sqrt{T}vec(\hat{\Phi}_n - \Phi_n) \rightarrow_d$

$N(0, G_n \Sigma G_n^T) \quad \forall n = 1, 2, \dots, \sqrt{T} \text{vec}(\hat{\Psi}_m - \Psi_m) \rightarrow_d N(0, F_m \Sigma F_m^T) \quad \forall m = 1, 2, \dots$ , where  $G_n = \partial \text{vec} \Phi_n / \partial \text{vec}(A_1, \dots, A_p)^T$  and  $F_m = G_1 + \dots + G_m$ . See Lütkepohl (1990) and Lütkepohl and Reimers (1992) for details on the derivation of the covariance of the impulse response function. The derivation of the covariance of impulse response functions is not a focus of the research in this defense.

**Extension: Confidence Intervals**

Below, we show two general computational methods based on simulation that can be used to determine asymptotics for impulse responses.

**Algorithm 2.** *Monte Carlo confidence interval (Sheppard, 2010)*

For a VAR model in companion form:  $y_t = Ay_{t-1} + \epsilon_t$

1. Compute  $\hat{A}$  from the initial data and estimate  $\hat{\Sigma}$  in the asymptotic distribution  $\sqrt{T}(\hat{A} - A) \tilde{N}(0, \Sigma)$
2. Use  $\hat{A}$  and  $\hat{\Sigma}$  to generate simulated values of  $\hat{A}_b$  from  $\sqrt{\hat{\Sigma}}\epsilon + \hat{A}$  where  $\epsilon \tilde{N}(0, 1)$
3. Use  $\hat{A}_b$  to compute and store impulse responses  $\hat{\Phi}_b \quad \forall b = 1, 2, \dots, B$ .
4. Goto 2 and compute  $B$  iterations
5. For each time, order the  $\hat{\Phi}_b \quad \forall b = 1, 2, \dots, B$ , the 5<sup>th</sup> and 95<sup>th</sup> percentile of the distribution are the confidence interval.

**Algorithm 3.** *Bootstrap confidence interval (Sheppard, 2010)*

For a VAR model in companion form:  $y_t = Ay_{t-1} + \epsilon_t$

1. Compute  $\hat{A}$  from the initial data and generate residuals  $\hat{\epsilon}_t$
2. Compute new set of residuals  $\epsilon_t^b$  from  $\epsilon_t$  by sampling with replacement
3. Use  $\hat{A}$  estimated coefficients and new residuals  $\epsilon_t^b$  to generate  $y_t^b$
4. Use  $y_t^b$  to re-estimate the coefficients  $\hat{A}^b \quad \forall b = 1, 2, \dots, B$  and compute  $\hat{\Phi}_b$
5. Goto 2 and compute  $B$  iterations
6. For each time, order the  $\hat{\Phi}_b \quad \forall b = 1, 2, \dots, B$ , the 5<sup>th</sup> and 95<sup>th</sup> percentile of the distribution are the confidence interval.

Both techniques can be used to determine confidence intervals from impulse response func-

## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

tions.

### 4.8 Empirical Implementation

We conduct empirical analysis on the 24 Korean companies using the VARX estimated through adaptive lasso. However, as per Tibshirani (1996), the regressors in lasso require to be standardized. The betas in the regression need to be comparable in magnitude so that the optimization procedure which aims to shrink the betas do not simply cut off the larger betas caused from the regressors being of different scales. In the VAR case, our  $Y = vec(Y^*)$  also require to be standardized, as our dependent vector consists of more than one variable. These variables need to be of a comparable size for adaptive lasso to work.

However, by standardizing our intraday factors, durations, quote revisions and trades, the beta coefficients produced are no longer as interpretable in application. For example, we cannot employ them to generate impulse response functions. Furthermore, they are no longer comparable with results produced by Hasbrouck (1991) or Dufour and Engle (2000). This can be resolved via a simple transformation of the estimated beta coefficients. It is clear to readers that in  $Y = X\beta + U$  when both  $Y$  and  $X$  are standardized, then there is no constant term in the model, as all the time series are centered around zero. Let  $\beta^{std}$  be the estimated standardized coefficients, then the raw coefficients can be determined as,

$$\beta_k = \beta_k^{std} \frac{s_y}{s_{x_k}}$$

where  $s_y$  and  $s_{x_k}$  refers to the sample standard deviation of the dependent variable and the  $k^{th}$  regressor variable.  $\beta_k^{std}$  related to quote revisions  $r_t$  is adjusted by  $\frac{s_r}{s_{x_k}}$  whilst  $\beta_k^{std}$  related to trades  $x_t$  is adjusted by  $\frac{s_x}{s_{x_k}}$ . Therefore, it is not difficult to convert adaptive lasso estimates derived under standardized data into unstandardized estimates comparable with results derived by Dufour and Engle (2000) without any further computation. In essence, we standardize our variables in order to have a fair adaptive lasso shrinkage, after which we 'un-standardize' our

estimates to make them comparable and interpretable. This ensures that our coefficients are consistent in magnitude and comparable to previous research, furthermore, we are able to generate impulse response functions where price impact results would be comparable to existing research.

Instead of estimating one set of adaptive lasso estimates for the full sample 2007-2012, we conduct multiple estimates at weekly frequency. We do this for two reasons. Firstly, the tick dataset is too large to conduct a single stage estimation for 6 years. Therefore, breaking the dataset into smaller manageable pieces is necessary. To put it in perspective, in Dufour and Engle's (2000) work, only 62 days are considered. Secondly, we are able to examine how price impact varies across time on a weekly frequencies. This becomes particularly useful for our research conducted in chapters 4 and 5.

## 4.9 Empirical Results

Here we present the estimated coefficients from the adaptive lasso VAR model. Table 4.1 provides average VAR coefficients for Samsung Electronics (005930 KS), the remaining estimated coefficients for the other 23 stocks are provided in appendix A.

Examining the coefficient signs and significance allows us to determine some relationships between these five microstructure variables. Below is our summary with regards to table 4.1, where we show the estimated coefficients for Samsung Electronics from 2007 to 2012. From the quote revision side of the VAR model we find:

These findings suggest some reversion in quote revisions, however this is not attributed to bid-ask bounce as we are using mid-point price for the calculation of quote revisions. Lagged trades have a decaying but significant positive contribution to quote revision. This is common sense, as it simply suggests a large buy (sell) volume is likely to push prices higher (lower). These endogenous relationships are entirely consistent with Dufour and Engle (2000), who document

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

**Table 4.1:** Summary on the VAR Quote Revision Equation

Factor	Lags	Impact on Quote Revisions
Quote revisions	1 - 10	significant negative contribution; this decreases in magnitude with increased lags
Signed log volume (trade)	0	significant positive contemporaneous contribution
Signed log volume (trade)	1 - 10	significant positive contribution; this decreases in magnitude with increased lags
Durations	0	significant positive contemporaneous contribution
Durations	1 - -	no significance
Spreads	0	significant positive contemporaneous contribution
Spreads	1 - 4	significant positive contribution
Depth	0	significant negative contemporaneous contribution
Depth	1 - 6	significant negative contribution

similar dynamics, despite only using signed trade indicators with no log volume information. However, the effect of our exogenous variables on quote revisions is more interesting.

Firstly, contrary to Dufour and Engle (2000), we show that most of the lagged trade durations do not have a significant impact and contemporaneous duration has a positive relationship with quote revisions, meaning that if it took a long time for the next trade to occur, then it is likely to have a stronger price impact. This is contrary to the "fast trading is informed trading" belief, but consistent with Grammig, Theissen and Wunche (2011). We argue that a longer duration is sign of greater illiquidity, and therefore more substantial quote revision fluctuations. Likewise, Grammig, Theissen and Wunche (2011) also find that active markets have smaller price impact. Grammig, Theissen and Wunche (2011) explain the findings using Parlour's (1996) *crowding out effect*. Parlour (1996) state that in low information asymmetry markets with decent market liquidity, the crowding out of limit orders by market orders by impatient participants would cause high trade intensity and lower trade durations. Therefore, lower price impact would be associated with lower trade durations; this is consistent with our findings.

Furthermore, spreads have a positive relationship with quote revisions. This again follows the the illiquidity argument - a wider bid-ask spread is a classic illiquidity indicator, and is likely to cause greater price impact. Trade conditional depth have a negative relationship with quote revisions. Obviously the greater the depth, the less likely it is for a trade to "eat-up" the levels in the orderbook, and so the negative relationship is sensible.

From the trade side of the VAR model we find:

In general lagged quote revisions have a negative impact on trades - if prices go up (down),

**Table 4.2:** Summary on the VAR Trade Equation

Factor	Lags	Impact on Trades (signed log volume)
Quote revisions	1	significant positive contribution
Quote revisions	2 - 10	significant negative contribution; this decreases in magnitude with increased lags
Signed volume (trade)	1 - 10	significant positive contribution; this decreases in magnitude with increased lags
Durations		no significance
Spreads		no significance
Depth	1 - 5	significant negative contribution; this decreases in magnitude with increased lags

it discourages further buying (selling) on a microstructure level. This is however not true for the first lag, where there is a significant positive relationship. The net effect of lagged quote revisions is still negative for trades. Lagged trades have a significant positive relationship with trades. This indicates autocorrelation in trading - buy (sell) trades are followed by more buy (sell) trades. It partly could be the result of electronic trading and broker algorithms splitting large orders into multiple smaller packets to trade in attempt to reduce price impact. We note that durations and spreads do not have any meaningful significance in determining future trades. However, depth does have a strong negative relationship. Greater depth in the opposite side of the orderbook suggests either there is a lot of liquidity in the market, or that there's simply a lot of limit order participants that hold opposing views to the market order, and are willing to trade as its counterparty. Therefore opposing side depth somewhat counters the effect of trade autocorrelation, which is not seen with durations and spreads.

To simplify, we tabulate our findings as the following,

**Table 4.3:** Summary on the Factors Influencing Quote Revisions and Trades

Endogenous Variable	Influencing Factors	Significant Lags	Relationship
Quote revisions	Quote revisions	1 - 10	-
Quote revisions	Trades	0 - 10	+
Quote revisions	Durations	0	+
Quote revisions	Spreads	0 - 4	+
Quote revisions	Depth	0 - 6	-
Trades	Quote revisions	1 - 10	-
Trades	Trades	1 - 10	+
Trades	Durations		
Trades	Spreads		
Trades	Depth	1 - 5	-

Extending from Samsung Electronics, we aggregated our estimated coefficients for all 24 Korean companies in our sample. The results from the aggregate show a similar picture.

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

**Table 4.4:** Average Coefficients for Samsung Electronics

We estimate the adaptive lasso VAR model in section 4.1 on high frequency tick data on a weekly basis. Below we present the average coefficients over the period Jan 2007 to Dec 2012 for Samsung Electronics (005930 KS). Adaptive lasso VAR models are estimated from Jan 2007 to Dec 2012 at weekly frequencies for all 24 stocks listed in Table 2.4.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value		lag	mean	stdev	T-stat	p-value	
Quote Revisions	1	<b>-0.097</b>	0.070	-24.55	0.000	***	1	<b>9.857</b>	4.605	37.87	0.000	***
	2	<b>-0.065</b>	0.064	-18.09	0.000	***	2	<b>-9.513</b>	3.974	-42.35	0.000	***
	3	<b>-0.039</b>	0.048	-14.39	0.000	***	3	<b>-5.350</b>	2.423	-39.06	0.000	***
	4	<b>-0.032</b>	0.050	-11.17	0.000	***	4	<b>-4.448</b>	1.905	-41.31	0.000	***
	5	<b>-0.023</b>	0.037	-10.79	0.000	***	5	<b>-3.157</b>	1.524	-36.65	0.000	***
	6	<b>-0.016</b>	0.032	-8.79	0.000	***	6	<b>-2.289</b>	1.329	-30.48	0.000	***
	7	<b>-0.012</b>	0.031	-7.12	0.000	***	7	<b>-1.621</b>	1.109	-25.87	0.000	***
	8	<b>-0.007</b>	0.018	-6.80	0.000	***	8	<b>-1.063</b>	0.966	-19.47	0.000	***
	9	<b>-0.007</b>	0.017	-6.92	0.000	***	9	<b>-0.599</b>	0.710	-14.91	0.000	***
	10	<b>-0.006</b>	0.018	-5.85	0.000	***	10	<b>-0.259</b>	0.437	-10.49	0.000	***
Trades	0	<b>8.92E-04</b>	5.56E-04	28.39	0.000	***	1	<b>0.210</b>	0.078	47.91	0.000	***
	1	<b>2.68E-04</b>	2.13E-04	22.28	0.000	***	2	<b>0.066</b>	0.031	37.54	0.000	***
	2	<b>2.17E-04</b>	1.44E-04	26.68	0.000	***	3	<b>0.066</b>	0.013	89.30	0.000	***
	3	<b>1.40E-04</b>	1.37E-04	18.03	0.000	***	4	<b>0.052</b>	0.011	83.74	0.000	***
	4	<b>1.06E-04</b>	1.22E-04	15.39	0.000	***	5	<b>0.043</b>	0.010	76.86	0.000	***
	5	<b>6.87E-05</b>	9.86E-05	12.33	0.000	***	6	<b>0.038</b>	0.010	69.44	0.000	***
	6	<b>4.73E-05</b>	1.01E-04	8.32	0.000	***	7	<b>0.033</b>	0.010	60.53	0.000	***
	7	<b>3.21E-05</b>	6.39E-05	8.90	0.000	***	8	<b>0.031</b>	0.009	60.17	0.000	***
	8	<b>2.54E-05</b>	6.49E-05	6.92	0.000	***	9	<b>0.028</b>	0.010	50.73	0.000	***
	9	<b>2.18E-05</b>	6.69E-05	5.77	0.000	***	10	<b>0.030</b>	0.009	57.92	0.000	***
Durations	0	<b>2.49E-06</b>	2.21E-06	19.94	0.000	***	1	5.22E-06	7.95E-05	1.16	0.247	
	1	-9.08E-08	4.70E-06	-0.34	0.733		2	-4.69E-06	9.02E-05	-0.92	0.358	
	2	-2.70E-07	5.56E-06	-0.86	0.391		3	2.07E-06	7.52E-05	0.49	0.627	
	3	8.48E-08	5.24E-06	0.29	0.775		4	2.35E-06	8.69E-05	0.48	0.633	
	4	2.36E-07	6.61E-06	0.63	0.528		5	2.65E-06	5.33E-05	0.88	0.379	
	5	-1.82E-08	2.70E-06	-0.12	0.905		6	-6.30E-07	2.84E-05	-0.39	0.695	
	6	-6.51E-09	2.96E-06	-0.04	0.969		7	-5.41E-07	4.68E-05	-0.20	0.838	
	7	-2.45E-08	2.85E-06	-0.15	0.879		8	-2.06E-06	6.18E-05	-0.59	0.557	
	8	-1.60E-07	3.21E-06	-0.88	0.378		9	1.04E-06	3.81E-05	0.48	0.630	
	9	-1.06E-08	2.82E-06	-0.07	0.947		10	-3.79E-06	4.12E-05	-1.62	0.105	
Spreads	0	<b>3.15E-06</b>	1.47E-05	3.80	0.000	***	1	-5.05E-06	3.88E-04	-0.23	0.818	
	1	<b>1.94E-06</b>	1.11E-05	3.08	0.002	***	2	-7.95E-07	1.42E-04	-0.10	0.921	
	2	<b>1.36E-06</b>	9.08E-06	2.65	0.009	***	3	-3.87E-06	1.30E-04	-0.53	0.599	
	3	1.21E-06	1.05E-05	2.04	0.042	**	4	6.02E-06	7.72E-05	1.38	0.168	
	4	1.09E-06	7.47E-06	2.57	0.011	**	5	9.88E-07	7.42E-05	0.24	0.814	
	5	-5.33E-08	6.22E-06	-0.15	0.880		6	8.07E-06	6.22E-05	2.30	0.022	**
	6	-1.79E-07	4.98E-06	-0.64	0.525		7	4.26E-06	4.08E-05	1.84	0.066	*
	7	3.41E-07	5.34E-06	1.13	0.259		8	-1.96E-06	4.24E-05	-0.82	0.415	
	8	1.53E-07	5.37E-06	0.51	0.614		9	2.28E-06	4.93E-05	0.82	0.415	
	9	2.39E-07	4.43E-06	0.96	0.339		10	1.18E-06	6.77E-05	0.31	0.759	
Depth	0	<b>-7.47E-07</b>	5.87E-07	-22.53	0.000	***	1	<b>-1.10E-05</b>	1.40E-05	-13.86	0.000	***
	1	<b>-2.21E-07</b>	3.47E-07	-11.24	0.000	***	2	<b>-3.97E-06</b>	8.80E-06	-8.00	0.000	***
	2	<b>-1.07E-07</b>	1.99E-07	-9.54	0.000	***	3	<b>-1.97E-06</b>	5.67E-06	-6.15	0.000	***
	3	<b>-7.33E-08</b>	1.52E-07	-8.50	0.000	***	4	<b>-1.24E-06</b>	5.95E-06	-3.69	0.000	***
	4	<b>-4.24E-08</b>	1.12E-07	-6.73	0.000	***	5	<b>-8.69E-07</b>	3.24E-06	-4.75	0.000	***
	5	<b>-2.90E-08</b>	1.32E-07	-3.89	0.000	***	6	-3.44E-07	2.76E-06	-2.20	0.028	**
	6	<b>-3.72E-08</b>	1.88E-07	-3.51	0.001	***	7	7.94E-08	2.68E-06	0.52	0.601	
	7	-9.27E-09	1.33E-07	-1.23	0.219		8	3.72E-07	3.66E-06	1.80	0.073	*
	8	-7.06E-09	1.14E-07	-1.10	0.273		9	2.79E-07	3.00E-06	1.64	0.102	
	9	-3.44E-09	8.55E-08	-0.71	0.477		10	5.97E-07	3.40E-06	1.11	0.272	

## 4.9 Empirical Results

**Table 4.5:** Average Coefficients for All Stocks

We estimate the adaptive lasso VAR model in section 4.1 on high frequency tick data on a weekly basis. Below we present the average coefficients over the period Jan 2007 to Dec 2012 for all 24 stocks listed in Table 2.4. We also present the percent of stocks in our sample that were significant for each coefficient.

	Quote Revision Equation					Trade Equation				
	lag	mean	% Significant			lag	mean	% Significant		
			1	5	10			1	5	10
Quote Revisions	1	-0.122	100%	100%	100%	1	10.902	100%	100%	100%
	2	-0.062	100%	100%	100%	2	-13.102	100%	100%	100%
	3	-0.038	100%	100%	100%	3	-5.580	100%	100%	100%
	4	-0.023	100%	100%	100%	4	-4.322	100%	100%	100%
	5	-0.014	100%	100%	100%	5	-2.674	100%	100%	100%
	6	-0.010	100%	100%	100%	6	-1.806	100%	100%	100%
	7	-0.006	100%	100%	100%	7	-1.153	96%	100%	100%
	8	-0.005	96%	96%	100%	8	-0.711	96%	100%	100%
	9	-0.003	79%	88%	88%	9	-0.386	96%	96%	96%
	10	-0.002	88%	92%	92%	10	-0.172	83%	92%	96%
Trades	0	1.52E-03	100%	100%	100%	1	0.268	100%	100%	100%
	1	3.00E-04	100%	100%	100%	2	0.069	100%	100%	100%
	2	2.32E-04	100%	100%	100%	3	0.071	100%	100%	100%
	3	1.05E-04	100%	100%	100%	4	0.048	100%	100%	100%
	4	6.36E-05	100%	100%	100%	5	0.040	100%	100%	100%
	5	4.00E-05	100%	100%	100%	6	0.033	100%	100%	100%
	6	2.47E-05	92%	100%	100%	7	0.028	100%	100%	100%
	7	1.71E-05	83%	88%	88%	8	0.026	100%	100%	100%
	8	1.16E-05	75%	83%	88%	9	0.024	100%	100%	100%
	9	8.36E-06	71%	79%	79%	10	0.027	100%	100%	100%
10	3.21E-06	50%	67%	67%						
Durations	0	4.67E-06	100%	100%	100%	1	1.49E-06	0%	0%	13%
	1	7.29E-08	4%	8%	13%	2	3.35E-06	8%	17%	25%
	2	3.40E-08	0%	0%	13%	3	2.48E-06	0%	4%	8%
	3	1.78E-08	0%	0%	0%	4	2.85E-06	4%	13%	29%
	4	6.12E-08	0%	4%	8%	5	1.70E-06	0%	4%	8%
	5	-9.19E-09	0%	8%	21%	6	1.32E-06	4%	8%	21%
	6	-5.10E-08	0%	0%	4%	7	1.51E-06	0%	8%	13%
	7	-1.20E-08	0%	4%	13%	8	3.42E-07	4%	13%	13%
	8	-9.68E-08	4%	8%	17%	9	8.95E-07	4%	8%	21%
	9	-6.35E-08	4%	4%	13%	10	-5.97E-07	0%	8%	13%
10	4.14E-09	0%	0%	0%						
Spreads	0	4.80E-05	100%	100%	100%	1	4.97E-04	71%	71%	75%
	1	1.94E-05	83%	92%	92%	2	1.52E-04	21%	33%	63%
	2	1.15E-05	50%	71%	79%	3	9.24E-05	13%	33%	50%
	3	8.41E-06	25%	46%	54%	4	9.62E-05	21%	33%	38%
	4	5.97E-06	4%	25%	38%	5	8.34E-05	13%	21%	33%
	5	2.01E-06	8%	13%	25%	6	9.05E-05	8%	25%	42%
	6	2.09E-06	0%	0%	8%	7	6.61E-05	4%	21%	33%
	7	6.47E-07	0%	4%	8%	8	5.54E-05	4%	17%	17%
	8	8.76E-07	0%	4%	13%	9	5.46E-05	4%	13%	17%
	9	1.54E-06	4%	25%	38%	10	5.21E-05	4%	29%	33%
10	4.17E-07	0%	25%	29%						
Depth	0	-1.74E-06	100%	100%	100%	1	-8.79E-06	100%	100%	100%
	1	-3.24E-07	100%	100%	100%	2	-1.01E-06	58%	75%	75%
	2	-1.55E-07	100%	100%	100%	3	-1.14E-07	50%	63%	75%
	3	-4.50E-08	75%	79%	79%	4	4.00E-07	21%	58%	63%
	4	-3.33E-08	33%	50%	58%	5	6.54E-07	38%	54%	75%
	5	-1.33E-08	17%	17%	21%	6	1.02E-06	67%	88%	92%
	6	-1.83E-09	8%	21%	29%	7	1.08E-06	75%	83%	88%
	7	-1.60E-09	4%	8%	13%	8	1.09E-06	71%	83%	92%
	8	3.62E-09	13%	13%	25%	9	1.24E-06	75%	83%	83%
	9	-6.26E-09	0%	13%	25%	10	1.41E-06	0%	0%	0%
10	4.74E-09	29%	38%	42%						



## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

### 4.10 Conclusions

In conclusion, we have thoroughly analyzed the role trading activity plays in price discovery. We have presented our new price impact model which incorporates adaptive lasso estimation. Our results show that aside from durations and trades, order book measures such as spreads and depths also have a significant role in affecting price impact. Price impact is greater when in periods where spreads are wider and depth is small. We show that price impact is not only driven by the information content of incoming trades but also by existing order book illiquidity.

### 4.11 Extensions on Price Impact

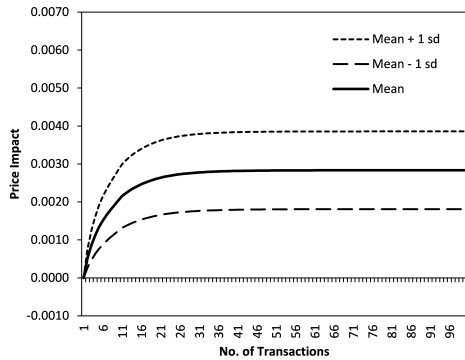
Our analysis so far has been concerned with the coefficients of the VARX model. These have shown the significance of the order book in impacting quote revisions. In this section, we account for the impact of order book illiquidity and estimate price impact in the spirit of Hasbrouck (1991). Following from the coefficients derived in the VARX model (such as those in table 4.4), we are able to construct cumulative impulse response functions to determine price impact. These estimates are not sensitive to order book illiquidity. In figure 4.1, we illustrate for Samsung Electronics, the CIRF to quote revisions due to a trade shock (i.e., price impact of a trade) and the CIRF to trades due to a trade shock (i.e., trade impact of a trade). We can see that price impact is almost completely realized after 30 transactions; this is similarly true for trade impact. In table 4.6, we tabulate the price impact for all 24 companies in our sample.

To further illustrate the variability of the CIRFs across time, below we plot both price impact and trade impact across time. In the following chapters, we will try and understand some of the time varying dynamics of price impact.

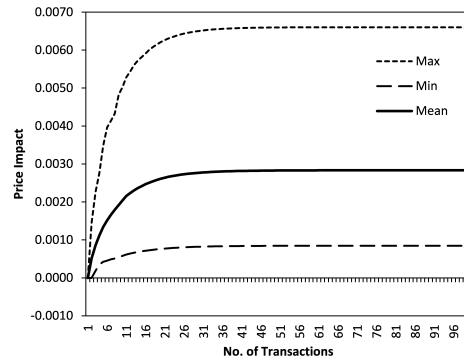
We document the average price impact of each stock in our sample. We find that stocks with higher average price impact also tend to have higher price impact variability across time. Therefore, stock with greater information content of trades (as our VARX model already accounts for order book illiquidity), also has greater variability in the levels of information content. In many ways, this makes sense. To borrow the concept of news in Easley and O'Hara (1992), we argue that in periods of no news, the information content of trades will be low, as very little private

**Figure 4.1:** Price Impact and Trade Impact: Cumulative Impulse Response Functions (CIRF)

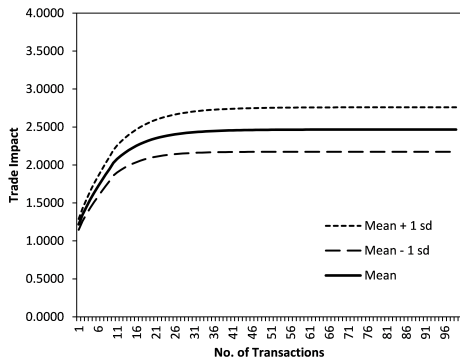
We plot the CIRFs of Samsung Electronics (005930 KS) derived from the estimates of our VARX model (see table 3.1). In plot (1,1) we plot the mean price impact across full sample history (January 2007 to December 2012) along with the 1 standard deviation bounds. In plot (1,2) we plot the mean price impact across full sample history (January 2007 to December 2012) along with the maximum and minimum price impact curves during the full sample history. In plot (2,1) we plot the mean trade impact across full sample history (January 2007 to December 2012) along with the 1 standard deviation bounds. In plot (2,2) we plot the mean trade impact across full sample history (January 2007 to December 2012) along with the maximum and minimum trade impact curves during the full sample history.



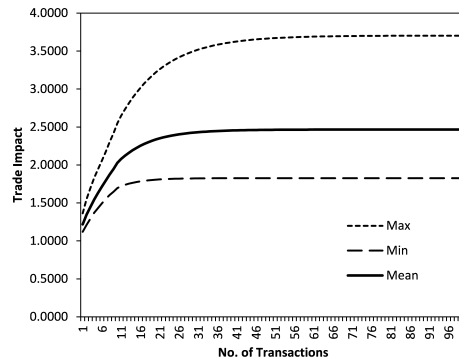
(a) CIRF (quote revision) +/- 1 sd



(b) CIRF (quote revision) min max



(c) CIRF (trades) +/- 1 sd



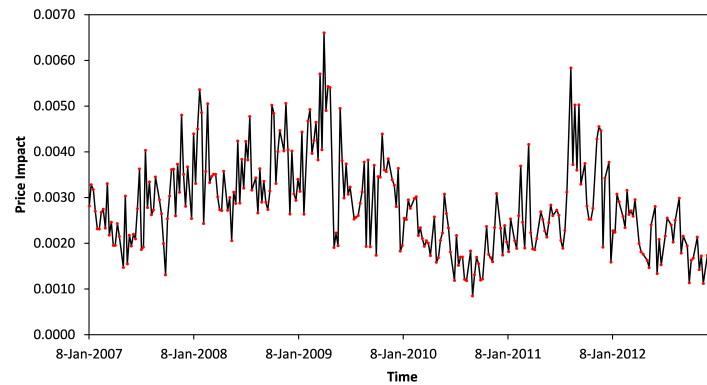
(d) CIRF (trades) min max

## 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

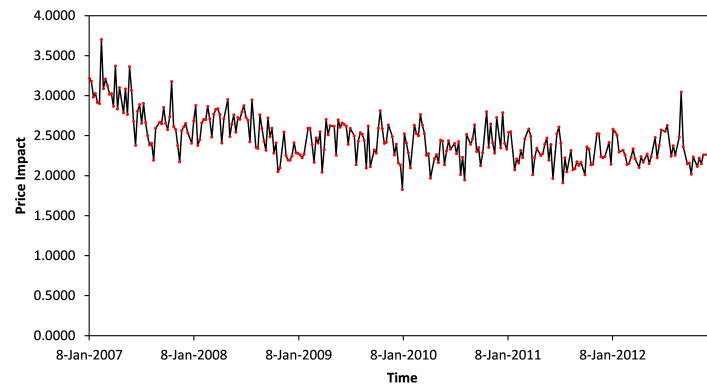
---

**Figure 4.2:** Price Impact and Trade Impact: Variability across Time

We plot the CIRFs of Samsung Electronics (005930 KS) across time. Each week, we run estimates of our VARX model and determine the CIRFs. Table 1 plots price impact and table 2 plots trade impact from January 2007 to December 2012.



(a) Price Impact of a Trade



(b) Trade Impact of a Trade

## 4.11 Extensions on Price Impact

information is held by informed traders. In periods where there is news, the information content of trades will increase as informed individuals will trade against liquidity providers for profit, and liquidity providers will adjust their quotes accordingly to incoming order flow information (see Kyle, 1985). Therefore, for stocks with little news, price impact will remain low with little variability and for stocks with a lot of news, price impact will increase whenever a news event occurs, this increases overall average price impact as well as increasing price impact variability.

**Table 4.6:** Average Price Impact for All Stocks

We present the mean and standard deviation of price impact for the period January 2007 to December 2012.

Ticker	Company	KOSPI 200 Rank	Mean	Std Dev	Max	Min
030200 KS	KT Corp	19	0.0021	0.0011	0.0073	0.0003
015760 KS	KEPCO	13	0.0022	0.0011	0.0074	0.0005
000660 KS	SK Hynix	6	0.0023	0.0009	0.0068	0.0009
055550 KS	Shinhan Group	5	0.0026	0.0010	0.0069	0.0004
005380 KS	Hyundai Motors	2	0.0027	0.0009	0.0064	0.0007
034220 KS	LG Display	22	0.0027	0.0011	0.0074	0.0008
005930 KS	Samsung Electronics	1	0.0028	0.0010	0.0066	0.0008
000270 KS	Kia Motors	7	0.0029	0.0014	0.0070	0.0007
066570 KS	LG Electronics	16	0.0029	0.0011	0.0067	0.0008
005490 KS	POSCO	3	0.0031	0.0012	0.0068	0.0009
017670 KS	SK Telecom	11	0.0031	0.0018	0.0104	0.0000
086790 KS	Hana Financial	18	0.0032	0.0014	0.0085	0.0005
010140 KS	Samsung Heavy Industries	24	0.0032	0.0009	0.0068	0.0009
009150 KS	Samsung Electro-Mechanics	27	0.0033	0.0013	0.0079	0.0007
033780 KS	KT&G	17	0.0033	0.0013	0.0090	0.0012
000720 KS	Hyundai Engineering & Construction	29	0.0035	0.0012	0.0086	0.0011
012330 KS	Hyundai Mobis	4	0.0038	0.0014	0.0102	0.0015
010950 KS	S-Oil	23	0.0039	0.0019	0.0128	0.0004
003550 KS	LG Corp	24	0.0039	0.0014	0.0101	0.0006
051910 KS	LG Chemicals	10	0.0041	0.0015	0.0119	0.0011
009540 KS	Hyundai Heavy Industries	15	0.0042	0.0016	0.0101	0.0010
035250 KS	Kangwonland	28	0.0050	0.0020	0.0117	0.0010
000830 KS	Samsung C&T	20	0.0052	0.0026	0.0155	0.0009
051900 KS	LG Household & Healthcare	26	0.0098	0.0042	0.0323	0.0014

In figure 4.3, we illustrate that stocks with higher price impact tend to also have higher price impact variability across time.

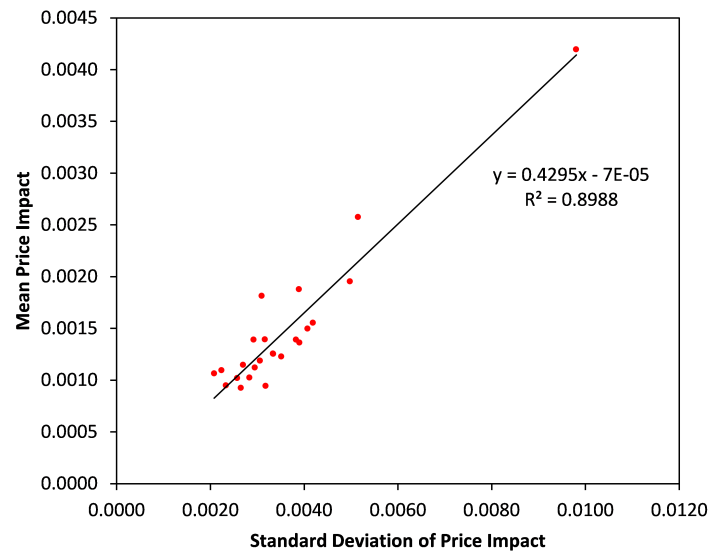
In figure 4.4 we illustrate cross-sectional variability and mean price impact across time. The black line in figure 4.4a plots the mean price impact of the 24 stocks from January 2007 to December 2012 and the red line in figure 4.4a plots the standard deviation of price impact of the 24 stocks across the same time period. We show that in periods where cross-sectional variability in price impact between stocks is high, the mean cross-sectional price impact across stocks is

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

**Figure 4.3:** Price Impact: Time-series Variability vs Mean

We plot the mean and standard deviation of price impact for all 24 stocks in our sample. We show visually that companies with higher mean price impact across time tend to also have higher price impact variability across time.



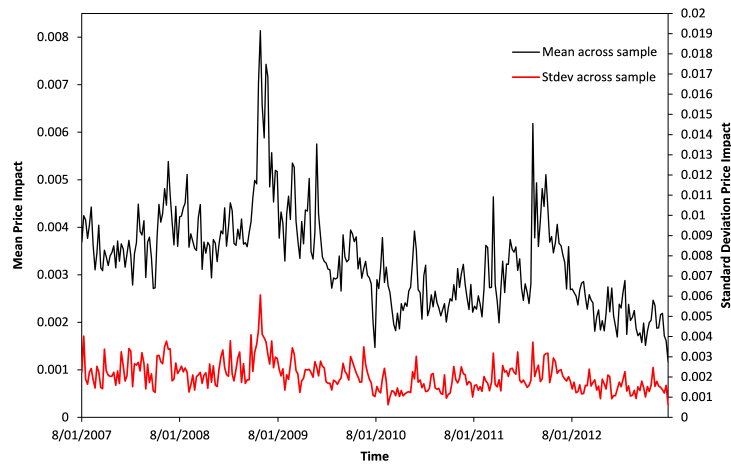
(a) Price Impact of a Trade

also high. This indicates that when there is market-wide news, different stocks have different impact levels hence causing greater dispersion in price impact.

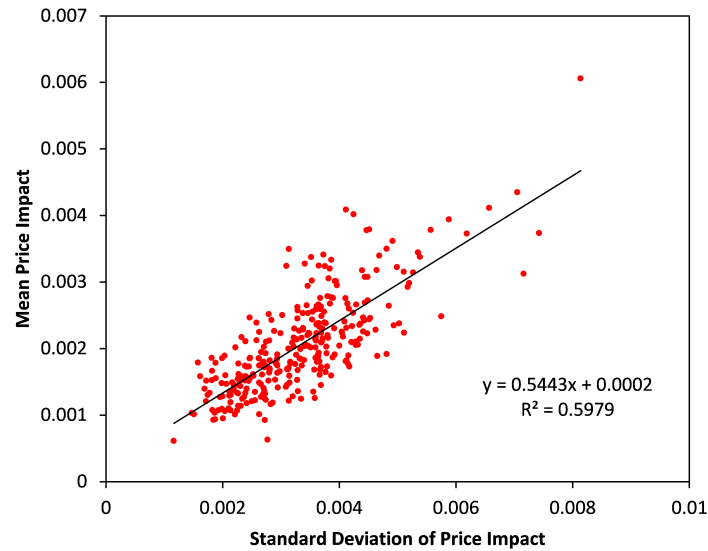
In this subsection we have shown some stylized facts on the variability of price impact. In chapters 5 and 6 we will provide greater detail in documenting the relationships between price impact and other trading metrics in time-series.

**Figure 4.4:** Price Impact: Cross-sectional Variability vs. Mean

We define cross-sectional variability to be the standard deviation of price impact across the 24 stocks in the sample at a particular point in time. Mean price impact refers to the average mean across the 24 stocks in the sample at a particular point in time. The graphs illustrate that in periods where price impact is high, the variability of price impact between stocks is also high. The peaks perhaps correspond to systematic news events that impact across all stocks, but at different levels, therefore increasing both the mean and the variability of price impact.



(a) Mean and Std. Dev. of Price Impact



(b) Mean Std. Dev. Price Impact Scatterplot

#### 4. THE IMPACT OF INFORMATION CONTENT AND ILLIQUIDITY ON QUOTE REVISIONS

---

# References

- [1] Biais, B., Hillion, P. and C. Spatt (1995) An empirical analysis of the limit order book and the order flow in the Paris Bourse, *Journal of Finance* 50, 1655 - 1689
- [2] Brownlees, C.T. and G.M. Gallo (2006) Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics & Data Analysis* 51, 2232 - 2245
- [3] Cao, C., Hansch, O. and X. Wang (2009) The information content of an open limit-order book, *Journal of Futures Markets* 29, 16 - 41
- [4] Chan, K., Chung, Y. and W. Fong (2002) The informational role of stock and option volume, *Review of Financial Studies* 15, 1049 - 1075
- [5] de Jong, F., Nijman, T. and A. Roell (1995) A comparison of the cost of trading French shares on the Paris Bourse and on SEAQ International, *European Economic Review* 39, 1277 - 1301
- [6] Diamond, D.W. and R.E. Verrecchia (1987) Constraints on short-selling and asset price adjustment to private information, *Journal of Financial Economics* 18, 277 - 311
- [7] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [8] Dufour, A. and R.F. Engle (2000a) The ACD model: predictability of the time between consecutive trades, *ISMA Centre, Discussion papers in Finance*, The University of Reading, UK



## REFERENCES

---

- [9] Easley, D. and M. O'Hara (1992) Time and the process of security price adjustment, *Journal of Finance* 47, 577 - 605
- [10] Engle, R.F. and A.J. Patton (2004) Impacts of trades in an error-correction model of quote prices, *Journal of Financial Markets* 7, 1 - 25
- [11] Engle, R.F. and J.R. Russell (1998) Autoregressive conditional duration: a new model for irregularly spaced transaction data, *Econometrica* 66, 1127 - 1162
- [12] Escribano, A. and R. Pascual (2006) Asymmetries in bid and ask responses to innovations in the trading process, *Empirical Economics* 30, 913 - 946
- [13] Fan, J. and R. Li (2001) Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties, *Journal of the American Statistical Association* 96 No. 456, 1348 -1360
- [14] Foster, F.D. and S. Viswanathan (1993) Variations in trading volume, return volatility and trading costs: evidence on recent price formation models, *Journal of Finance* 48, 187 - 211
- [15] Grammig, J., Theissen, E. and O. Wünsche (2011) Time and the price impact of a trade: a structural approach, working paper, University of Tübingen, Germany
- [16] Glosten, L. and L. H. Harris (1988) Estimating the components of the bid-ask spread, *Journal of Financial Economics* 21, 123 - 142
- [17] Glosten, L. and P. Milgrom (1985) Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71 - 100
- [18] Goldstein, M.A. and K.A. Kavajecz (2000) Eighths, sixteenth, and market depth: changes in tick size and liquidity provision on the NYSE, *Journal of Financial Economics* 56, 125 - 149
- [19] Hafner, C.M. (2005) Durations, volume and the prediction of financial returns in transaction time, *Quantitative Finance* 5 No.2, 145 - 152
- [20] Harris, L. and V. Panchapagesan (2005) The informational content of the limit order book: evidence from NYSE specialist trading decisions, *Journal of Financial Markets* 8, 25 - 67

- 
- [21] Hasbrouck, J. (1991) Measuring the information content of stock trades, *Journal of Finance* 46, 179 - 208
- [22] Hasbrouck, J. (1991a) The summary informativeness of stock trades: an econometric analysis, *Review of Financial Studies* 4 No.3, 571 - 595
- [23] Hastie, T. and B. Efron (2011) Package 'lars', *CRAN repository*
- [24] Haustch, N. and R.H. Huang (2011) The market impact of a limit order, *Journal of Economic Dynamics and Control*, forthcoming
- [25] Hendershott, T., Jones, C.M. and A.J. Menkveld (2011) Does algorithmic trading improve liquidity?, *Journal of Finance* 66, 1 - 33
- [26] Hsu, N.J., Hung, H.L. and Y.M. Chang (2008) Subset selection for vector autoregressive process using Lasso, *Computational Statistics and Data Analysis* 52, 3645 - 3657
- [27] Huang, R.D. and H.R. Stoll (1997) The components of the bid-ask spread: a general approach, *Review of Financial Studies* 10, 995 - 1034
- [28] Jang, H. and P.C. Venkatesh (1991) Consistency between predicted and actual bid-ask quote revisions, *Journal of Finance* 46, 433 - 446
- [29] Kraemer, N., Schaefer, J. and A.L. Boulesteix (2009) Regularized estimation of large-scale gene regulatory networks with gaussian graphical models, *BMC Bioinformatics*, 10:384
- [30] Kraemer, N. and J. Schaefer (2011) Package 'parcor', *CRAN repository*
- [31] Kyle, A.S. (1985) Continuous auctions and insider trading, *Econometrica* 53 - 6, 1315 - 1336
- [32] Lee, C. and M. Ready (1991) Inferring trade direction from intraday data, *Journal of Finance* 46, 733 - 746
- [33] Lo, I. and S.G. Sapp (2006) A structural error-correction model of best prices and depth in the foreign exchange limit order market, *Documents de travail de la Banque de Canada*

## REFERENCES

---

- [34] Madhavan, A., Richardson, M. and M. Roomans (1997) Why do security prices change? A transaction-level analysis of NYSE stocks, *Review of Financial Studies* 10, 1035 - 1064
- [35] Manganello, S. (2005) Duration, volume and volatility impact of trades, *Journal of Financial Markets* 8, 377 - 399
- [36] Næs, R. and J.A. Skjeltorp (2006) Order book characteristics and the volume-volatility relation: empirical evidence from a limit order market, *Journal of Financial Markets* 9, 408 - 432
- [37] Pascual, R., Escibano, A. and M. Tapia (2004) Adverse selection costs, trading activity and price discovery in the NYSE: An empirical analysis *Journal of Banking & Finance* 28, 107 - 128
- [38] Pötscher, B.M. and U. Schneider (2009) On the distribution of the adaptive Lasso estimator, *Journal of Statistical Planning and Inference* 139, 2775 - 2790
- [39] Pesaran, H.H. and Y. Shin (1998) Generalized impulse response analysis in linear multivariate models, *Economics Letters* 58, 17 - 29
- [40] Ren, Y.W. and X.S. Zhang (2010) Subset selection for vector autoregressive processes via adaptive Lasso, *Statistics and Probability Letters* 80, 1705 - 1712
- [41] Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B* 58-1, 267-288
- [42] Tombeur, G. and G. Wuyts (2011) Does information in the limit order book help to predict returns?, working paper, Katholieke Universiteit Leuven, Belgium
- [43] Wu, Z. (2012) On the intraday periodicity duration adjustment of high-frequency data, *Journal of Empirical Finance* 19, 282 -291
- [44] Zou, H. (2006) The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101 No.476, 1418 - 1429

## REFERENCES

---

- [45] Van Ness, B.F., Van Ness, R.A. and R.S. Warr (2002) Is the adverse selection component really higher on the NYSE/Amex than on the Nasdaq? *Journal of Business Finance and Accounting* 29 807 - 824

## REFERENCES

---

## 5

# The Price Impact of a Trade and its linkage with Volatility

1

**Abstract:** *This chapter provides a note on the linkage between price impact and volatility. Price impact measures the information content of a trade (Hasbrouck, 1991; Dufour and Engle, 2000). Ross (1989) shows that volatility is positively correlated with information arrival. Price impact is also a measure for risk aversion (Pascual, Escribano and Tapia, 2004) and risk aversion is higher in periods of uncertainty, marked by greater volatility. Therefore, we hypothesize that price impact should have a positive relationship with volatility. Firstly, we explain that price impact as per Hasbrouck (1991) provides a good empirical proxy for Kyle's  $\lambda$  (Kyle, 1985). Using a simple single-period rational expectations equilibrium information model, we show that the size of Kyle's  $\lambda$  is directly and positively related to information flow volatility; and therefore expect price impact to be positively correlated to volatility. Secondly, we conduct time-varying price impact from January 2007 to December 2012 using our VARX specifications in chapter 4 and confirm its relationship with volatility. We show that price impact is indeed strongly related to volatility validating the results of our information model.*

---

<sup>1</sup>This chapter is directly linked to chapter 4. Some details on the construction of price impact is omitted as it has already been discussed in chapter 4.

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

### 5.1 Introduction

Whilst there has been a plethora of empirical market microstructure literature relating to trading design and its associated frictions, few works have provided linkage between market microstructure and asset pricing. Here we focus on studying the linkage between price impact (an important component in market microstructure) and volatility (an important component in asset pricing), both empirically and theoretically. We ask whether market microstructure interactions on a tick by tick basis are related to overall asset price volatility in longer frequencies. Daily prices in financial markets are formed from thousands of tick-by-tick transaction data, and it is therefore not unrealistic to assume that the behavior of these transactions will be related to longer term volatility characteristics. To the best of our knowledge, no existing research use time-varying impact to test its interactions volatility.

It has often been discussed that volatility is related to information flow. For example, in Clark (1973), volatility is used as a subordinator for proxying the speed of incoming information. Ross (1989) shows that price volatility is positively correlated with information arrival. Furthermore, Andersen (1996) and Andersen and Bollerslev (1997) relate information arrival to stochastic volatility and show that higher volatility is associated with the arrival of information. Since our price impact derived in chapter 4 is a measure of information content, we hypothesize a linkage between price impact and volatility. Higher price impact should relate to higher volatility, as both correspond to periods of high information content.

From another perspective, we also know that price impact is positively related to illiquidity. In chapter 4 and also in Grammig, Theissen and Wunsche (2011), we find that durations have a positive impact on quote revisions and subsequently price impact. Grammig, Theissen and Wunsche (2011) show that in periods of illiquidity as measured through longer durations, price impact is higher. Illiquidity is also known to be positively related to volatility, which is clearly shown in market microstructure inventory models of Stoll (1978), Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), Copeland and Galai (1983) and Foster and Viswanathan

(1990). Moreover, Pastor and Stambaugh (2003) document that the empirical correlation between aggregate liquidity and market volatility is negative. Therefore, the existing literature on liquidity and volatility also indicates a positive relationship between price impact and volatility.

Also, price impact is a measure of risk aversion (Escribano, Pascual and Tapia, 2004); and we are aware that risk aversion is higher in periods where greater uncertainty exists. Uncertainty can be measured through volatility. This also points to a positive relationship between price impact and volatility.

Therefore, (1) the *information flow - volatility* stream of literature, (2) the *liquidity - volatility* stream of literature and (3) the *risk aversion - volatility* argument all point to the possibility of a positive relationship between price impact and volatility. In this chapter, we validate this hypothesis.

In the remaining chapter, we construct a theoretical argument explaining why the positive relationship between price impact and volatility holds and subsequently document empirical evidence to uphold the claim. Our theoretical argument is based on (1) a rational expectation equilibrium model first suggested by Grossman (1976) and modified by Baker and Stein (2004) to include Kyle (1985) dynamics and (2) the fact that price impact is an empirical proxy to Kyle's  $\lambda$ . Our results show that Kyle's  $\lambda$  is positively related to informational volatility and inversely related to supply volatility. We validate this empirically by regressing price impact against several volatility measures and determine that it is indeed positively related.

In section 5.2 we show why price impact is an empirical proxy for Kyle's  $\lambda$ . In section 5.3 we construct a single period rational expectation equilibrium model explaining the relationship between price impact and volatility, section 5.4 we explain the empirical model we use to test the relationship, section 5.5 provides empirical results and section 5.6 concludes.



## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

### 5.2 Price Impact and Kyle's $\lambda$

Firstly, we show that price impact is an empirical proxy for estimating Kyle's (1985) illiquidity parameter  $\lambda$ .

Price impact (Hasbrouck, 1991; Dufour and Engle, 2000) measures the information content of trading or the risk aversion cost of trading against an insider (Pascual, Escribano and Tapia, 2004). From Hasbrouck's VAR model, price impact is derived from the cumulative quote revisions due to an incoming trade shock (i.e., how price changes due to incoming order flow). In chapter 4, we show that quote revisions are not only driven by the information content of incoming order flow as is suggested by Hasbrouck (1991), but also by order book illiquidity, such as the spread and depth just prior to the order. By constructing a VARX model, with three exogenous variables - durations, spreads and depth, we are able to account for these effects and measure price impact more accurately without the effects of order book illiquidity. In other words, our price impact estimation isolates the permanent cumulative quote revisions due to incoming order flow irrespective of levels of order book illiquidity.

Kyle (1985) considers a Bayesian-Nash equilibrium model for trading where the informed trader submits a market order that tries to maximize profit and the market maker uses information on the order flow to adjust prices so that market efficiency holds. From this model we note that,

$$P \propto \lambda(u + x) \tag{5.1}$$

where  $P$  is price,  $u$  is uninformed order flow,  $x$  is informed order flow and  $\lambda$  is the famous Kyle's lambda. Put simply, Kyle's  $\lambda$  is an illiquidity measure that captures how much price moves with the order flow. In Kyle's (1985) paper, it is also considered to be a way for the market maker to protect himself/herself from losing money to an informed trader.

Simple statistical measures for Kyle's  $\lambda$  include estimating the slope of the regression between absolute returns (dependent variable) and volume (independent variable). For short periods, it

is approximated as,

$$\hat{\lambda} = \frac{|\Delta P_t|}{Vol_t} \quad (5.2)$$

where  $P$  is price and  $Vol_t$  is volume. However, this measure is inaccurate as volume is a different concept to order flow. Simple linear regression taken directly from Kyle (1985), i.e.,  $\Delta P_t = \mu + \lambda x_t$ , is also problematic. For instance, Hasbrouck (1991) and Brennan and Subrahmanyam (1996) show that there is a lagged response between a trade and its impact on quote revisions. Therefore, market impact is not contemporaneous. If one tries to eliminate the lagged impact by implementing the regression at a longer frequency, then one would have to (1) justify a correct frequency and (2) show that information is not lost (and noise not accumulated) with the loss of granularity. This is troublesome.

Here we suggest price impact, as measured through the cumulative impulse response function (CIRF) of our VARX model, is a good empirical proxy for Kyle's (1985)  $\lambda$ . It utilizes tick data and captures fully the market impact of a trade. Firstly, they both describe the level of price movement due to order flow. Price impact examines average market impact at the tick level. Kyle's  $\lambda$  considers the market impact from a block of trades. In Kyle's theoretical model, the relationship between price and order flow is contemporaneous. In Hasbrouck (1991) and Brennan and Subrahmanyam (1996) (refer to the Hasbrouck-Foster-Viswanathan model), it is shown empirically that prices are not only impacted by contemporaneous order flow but also lagged order flow. Kyle's model is stylized and at a lower frequency than trade-by-trade. Empirically, estimating a contemporaneous relationship such as  $\Delta P_t = \mu + \lambda x_t$  where  $x_t$  is order flow is not enough to estimate the full magnitude of price impact. This is confirmed in the literature. In Pascual, Escribano and Tapia (2004) and Van Ness, Van Ness and Warr (2001), it is shown that Hasbrouck's reduced form approach to determine market impact is superior to structural models based on Glosten and Harris (1988) and Huang and Stoll (1997). In de Jong, Nijman and Roëll (1996), they show that VAR models captured twice as much price impact to the structural Glosten (1994) model for the Paris Bourse. This is because structural models assume that price impact from a trade is instantaneous, whilst the VAR model accounts for possible lags. Due to market frictions, it is shown that price impact is not immediate. Therefore, we suggest that

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

Kyle's  $\lambda$  is best estimated via the price impact model suggested in Hasbrouck (1991) and Dufour and Engle (2000).

### 5.3 A Single Period Rational Expectations Equilibrium Model

From the previous section, we argue that price impact is synonymous to Kyle's  $\lambda$  by definition. Therefore, here we construct a simple rational expectations equilibrium model that provides the relationship between Kyle's  $\lambda$  and volatility. Subsequently, price impact and volatility would follow with the same relationship. We show that Kyle's  $\lambda$  is positively related to informational volatility and inversely related to supply volatility. Our model is a simple single period model motivated from existing rational expectations equilibrium models of Grossman (1976), Harrison and Kreps (1978), Grossman and Stiglitz (1980), Baker and Stein (2004) and Hong, Scheinkman and Xiong (2006). In particular, our model utilizes the framework by Baker and Stein (2004).

Our initial setup is based on Grossman (1976). Let us consider the time period  $t = 0$  to 1 where there exists one risky asset and one risk-free asset. Let the risk-free asset be worth 1 dollar and the risky asset be worth  $P_0$  at time  $t = 0$ . The wealth at time  $t = 0$  is therefore  $W_0 = X_f + P_0X$ , where  $X_f$  and  $X$  are the number of units in the risk-free and risky asset respectively. The wealth at time  $t = 1$  is  $W_1 = (1 + r^f)X_f + P_1X$  where  $r^f$  is the risk-free rate. Substituting the time  $t = 0$  budget constraint into the wealth equation in  $t = 1$  yields,

$$W_1 = (1 + r^f)W_0 + (P_1 - P_0(1 + r^f))X \quad (5.3)$$

We assume that investors have constant absolute risk aversion (CARA) governed by a negative exponential utility function with a coefficient for absolute risk aversion denoted by  $a$  where  $a > 0$ .

$$U(W_1) = -e^{-aW_1} \quad (5.4)$$

$W_1$  is assumed to be normally distributed conditional on the information set  $I_0$ . Since the

### 5.3 A Single Period Rational Expectations Equilibrium Model

---

moment generating function of a Gaussian normal is  $M_X(s) = E(e^{sX}) = e^{\mu s + \frac{\sigma^2 s^2}{2}}$ . We substitute  $s$  for the absolute risk aversion coefficient  $a$ , and from which we obtain the expected conditional utility.

$$E(U(W_1)|I_0) = -e^{-aE(W_1|I_0) + \frac{a^2}{2}Var(W_1|I_0)} \quad (5.5)$$

It follows that the investor problem at time  $t = 0$  is equivalent to maximizing,

$$aE(W_1|I_0) - \frac{a^2}{2}Var(W_1|I_0) \quad (5.6)$$

Given that the investors can choose only between a risk-free asset and a risky stock at quantity  $X$ , solving the first order condition leads to the optimal individual demand of the risky stock to be,

$$X = \frac{E(P_1|I_0) - (1 + r^f)P_0}{aVar(P_1|I_0)} \quad (5.7)$$

This is a typical demand function governed by CARA. In our study, without a loss of granularity, let  $r^f = 0$ ,  $\gamma = \frac{1}{aVar(P_1|I_0)}$ ,  $V_0 = E(P_1|I_0)$  and  $V_1 = E(P_2|I_1)$ . From which,  $P_0 = V_0 - \frac{Q}{\gamma}$  and  $P_1 = V_1 - \frac{Q}{\gamma}$ . Therefore returns is simply the change in beliefs in valuation as the information set changes,  $\Delta P_1 = V_1 - V_0$ .

Since rational expectation revisions  $\Delta P_1$  is linear to net order flow  $f_1$  we write,

$$\Delta P_1 = \lambda f_1 \quad (5.8)$$

This formulation is similar to Kyle (1985) and Baker and Stein (2004).  $\lambda$  is equivalent to Kyle's illiquidity measure, it is constrained such that  $\lambda > 0$ . The net order flow  $f_1$  is decomposed into insider order flow  $m_1$  and liquidity order flow  $z_1$ , so that  $f_1 = m_1 + z_1$ . Furthermore, we expect the expected net order flow from liquidity traders to be zero, i.e.,  $E(z_1) = 0$ . We expect insider order flow to be positive,  $m_1 > 0$ , if the underlying change in information is positive, and vice versa if it is negative.

Let the true valuation for the risky asset at  $t = 0$  be  $F$ , and let the terminal value revealed via public announcement be  $F + \eta + \varepsilon$  at  $t = 2$ , where  $\eta$  is the value of the new information and

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

$\varepsilon$  accounts for some stochastic noise governed by a standard normal  $N(0, 1)$ . The insider tries to maximize his profits at  $t = 1$  as he alone is aware of the information. This is a single period insider maximization problem. In a single period model we do not need to consider monopolistic insiders that may hold information and trade strategically across multiple periods.

The insider's problem is (similar to Kyle, 1985),

$$\max E\{m_1(F + \eta + \varepsilon - P_1)\} \rightarrow \max E\{m_1(\eta - \Delta P_1)\} \quad (5.9)$$

where  $m_1$  is her order size. The insider trades by exploiting her private information  $\eta$  against the adverse price impact of trade  $\Delta P_1$ . Therefore,

$$\max E\{m_1(\eta - \lambda(m_1 + z_1))\} \quad (5.10)$$

Solving the first order conditions yields,

$$m_1 = \frac{\eta}{2\lambda} \quad (5.11)$$

which is the optimal insider's orderflow. The second derivative yields  $-2\lambda$  which proves we have solved the maximum.

Since  $\lambda = \frac{\text{cov}(\eta, f_1)}{\text{var}(f_1)} = \frac{\text{cov}(\eta, m_1 + z_1)}{\text{var}(m_1 + z_1)}$  (see Baker and Stein, 2004), we have

$$\beta = \frac{E(\eta(\frac{\eta}{2\lambda} + z_1)) - E(\eta)E(\frac{\eta}{2\lambda} + z_1)}{\text{var}(\frac{\eta}{2\lambda} + z_1)} \quad (5.12)$$

which is simplified as<sup>1</sup>,

$$\beta = \frac{\frac{1}{2\lambda} \text{var}(\eta)}{\frac{1}{4\lambda^2} \text{var}(\eta) + \text{var}(z_1)} \quad (5.13)$$

which can be expressed in quadratic form, but noting  $\lambda > 0$  constraint needs to be fulfilled, the equilibrium is,

$$\lambda = \sqrt{\frac{\text{var}(\eta)}{4\text{var}(z_1)}} \quad (5.14)$$

---

<sup>1</sup> The numerator simplifies to  $E(\frac{\eta^2}{2\lambda}) - E(\eta)E(\frac{\eta}{2\lambda}) = \frac{1}{2\lambda} \text{Var}(\eta)$  due to  $E(z_1) = 0$ . For the denominator, we note that  $z_1$  is independent to  $\eta$ , and therefore  $\text{Var}(\frac{\eta}{2\lambda} + z_1) = \frac{1}{4\lambda^2} \text{Var}(\eta) + \text{Var}(z_1)$ .

This result suggests that  $\lambda$  is driven by fluctuations in information  $\eta$  and fluctuations in liquidity supply  $z_1$ .

**Proposition 1.** The illiquidity parameter  $\lambda$  is directly and positively related to informational variance  $var(\eta)$ . The illiquidity parameter  $\lambda$  is inversely related to liquidity variance  $var(z_1)$ . Since  $var(z_1)$  can be assumed to be constant through time<sup>1</sup>,  $\lambda \propto var(\eta)$ . In periods where there is high volatility of the underlying information stream on a particular stock, then its Kyle's  $\lambda$  parameter will be high, subsequently price impact will be high.

Price impact should be positively related to informational variance. The result is not surprising, and is consistent with the arguments in Hasbrouck (1991) and Dufour and Engle (2000). Both papers suggested that price impact is a measure of informational content, which clearly indicates that price impact should be related to information flow and information variability. If we believe that return volatility is driven mostly by informational variance, then we would expect price impact to be positively correlated to volatility. We test this claim empirically in the sections below.

## 5.4 Empirical Tests

The dataset we use for this research comes from Thomson Reuters Tick History and records trade and quote data from the Korea Exchange (KRX) to the microsecond from January 2007 to December 2012. We pick 24 large capitalization stocks listed on the KRX with complete history over the specified time period. High frequency anomalies are cleaned using the Brownlees and Gallo (2006) approach and trade initiation is determined using Ellis, Michaely and O'Hara's (2000) rule. This rule supersedes the tick rule and Lee and Ready's (1991) algorithm in terms of accuracy. Details on the dataset and the methodologies we employ have stated in chapter 2. Price impact calculations are conducted using the adaptive lasso VARX methodology in chapter

---

<sup>1</sup>In most models, liquidity variance is considered as a constant (deterministic), such as Kyle (1985).

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

4.

In our initial probe, we plot the scatterplots between price impact and return volatility for Samsung Electronics. In the figure below, we consider three different measures for volatility at a weekly frequency<sup>1</sup>. We calculate price impact (using the VARX framework in chapter 4) by estimating model parameters using weekly tick data blocks. Hence we are able to obtain weekly frequency time-varying price impact. The figure shows a clear positive relationship between price impact and volatility, which is robust to all 3 volatility measures.

Here we conduct a simple regression analysis to determine the relationship between price impact  $\pi_t$  and volatility  $\sigma_t$ .

$$\sigma_t = \beta_0 + \beta_1 \pi_t + \varepsilon_t \quad (5.15)$$

This is conducted at a weekly frequency. Price impact is estimated from the cumulative impulse response function of a trade shock on quote revisions from the VARX model discussed in chapter 3. The parameters of the VARX model is estimated using weekly blocks of high frequency tick data.

We use three different proxies for the volatility measure  $\sigma_t$ ,

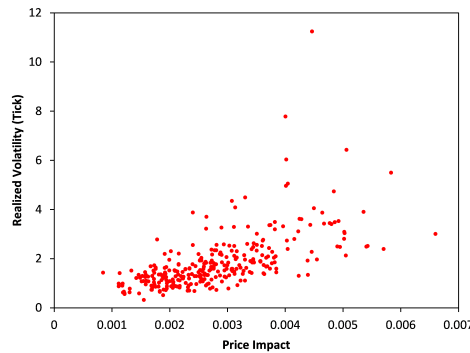
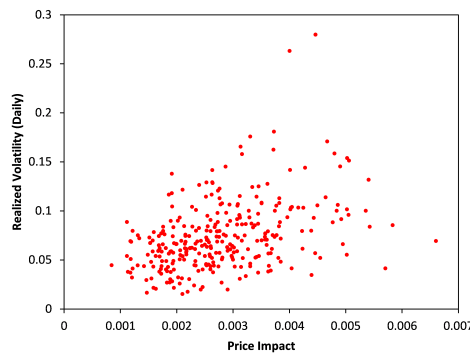
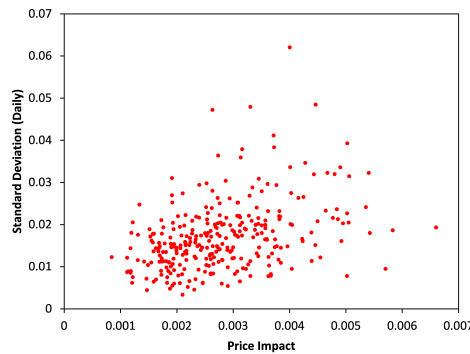
1. The standard deviation of daily asset returns over the week  $\sigma_t = \sqrt{\frac{1}{5} \sum_{i=1}^5 (R_i - \bar{R})^2}$
2. The realized volatility of daily asset returns over the week  $\sigma_t = \sum_{i=1}^5 |R_i|$
3. The realized volatility of quote revisions (changes in mid-price quotes, i.e., tick returns) over the week  $\sigma_t = \sum_{i=2}^T |\log(q_i) - \log(q_{i-1})|$  where  $T$  is the total number of transactions in the week and  $q_i$  is the midpoint price between the best bid and ask. Using the midpoint price bypasses fluctuations due to the bid-ask bounce.

---

<sup>1</sup>The definitions of the three volatility measures are provided later in this section

**Figure 5.1:** Price Impact and Volatility: Samsung Electronics Scatterplot

We plot the scatterplot between price impact as derived from our model in chapter 3 and several volatility measures. Realized volatility (tick) refers to the realized volatility of quote revisions (changes in mid-price quotes, i.e., tick returns) over the week  $\sigma_t = \sum_{i=2}^T |\log(q_i) - \log(q_{i-1})|$  where  $T$  is the total number of transactions in the week and  $q_i$  is the midpoint price between the best bid and ask. The realized volatility (daily) is of daily asset returns over the week  $\sigma_t = \sum_{i=1}^5 |R_i|$ , and standard deviation (daily) refers to the simple standard deviation metric of daily returns. The scatterplots use weekly data points from January 2007 to December 2012.

**(a)** Price Impact and Realized Volatility (Tick)**(b)** Trade Impact and Realized Volatility (Daily)**(c)** Trade Impact and Standard Deviation (Daily)



## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

### 5.5 Empirical Results

We document results from our regression in the tables below. Irrespective of the volatility measure, all 24 stocks on the Korea Exchange exhibited a strong positive relationship between price impact and volatility significant at the 1% level. Our empirical findings back our theoretical result that price impact is related to volatility. Furthermore, the positive relationship suggests that that return volatility is driven mostly by information variability. This finding also backs existing literature from Ross (1989) and Andersen (1996) that claim volatility is related to information.

Table 5.1: Linear Model: The Relationship between Volatility and Price Impact

This table documents the coefficients of a simple linear model on the relationship between volatility and price impact (liquidity). As per hypothesis 2, we expect a positive and statistically significant relationship between the price impact of trades and stock market volatility. Here we examine three different proxies for volatility denoted  $\sigma^2$ : (a) the standard deviation of asset returns (b) the realized volatility using absolute daily returns and (c) the realized volatility using absolute high frequency transaction returns (tick returns). The measure of price impact  $\pi$  is as per our definition on eqn 24. The regression runs from Jan 2007 to Dec 2012 at a weekly frequency.

$$\sigma_t^2 = \beta_0 + \beta_1 \pi_t + \varepsilon_t$$

	(a) Standard deviation with daily returns			(b) Realized volatility with absolute daily returns			(c) Realized volatility with absolute tick returns		
	$\beta_0$	$\beta_1$	Adj. $R^2$	$\beta_0$	$\beta_1$	Adj. $R^2$	$\beta_0$	$\beta_1$	Adj. $R^2$
005380KS	coeff 0.001	<b>20.452</b>	*** 27.4%	0.008	<b>86.690</b>	*** 25.9%	-1.481	<b>4,184.920</b>	*** 41.4%
	T-stat 0.766	10.743		0.907	10.327		-5.203	14.663	
	p-value 0.444	0.000		0.365	0.000		0.000	0.000	
005490KS	coeff 0.007	<b>7.822</b>	*** 17.1%	<b>0.026</b>	<b>37.110</b>	*** 17.8%	<b>-0.642</b>	<b>1,994.348</b>	*** 24.6%
	T-stat 4.653	8.000		3.924	8.191		-2.193	10.005	
	p-value 0.000	0.000		0.000	0.000		0.029	0.000	
005930KS	coeff 0.008	<b>7.756</b>	*** 15.7%	<b>0.033</b>	<b>35.963</b>	*** 17.6%	<b>-0.258</b>	<b>1,848.132</b>	*** 45.2%
	T-stat 6.676	7.585		6.254	8.130		-1.834	15.872	
	p-value 0.000	0.000		0.000	0.000		0.068	0.000	
012330KS	coeff 0.005	<b>10.152</b>	*** 22.2%	<b>0.024</b>	<b>41.680</b>	*** 21.2%	<b>-0.909</b>	<b>2,252.988</b>	*** 48.0%
	T-stat 3.120	9.424		3.393	9.139		-4.299	16.851	
	p-value 0.002	0.000		0.001	0.000		0.000	0.000	
055550KS	coeff 0.004	<b>16.481</b>	*** 30.6%	<b>0.015</b>	<b>72.259</b>	*** 30.7%	<b>-1.499</b>	<b>4,399.165</b>	*** 52.6%
	T-stat 2.565	11.735		2.302	11.749		-6.123	18.551	
	p-value 0.011	0.000		0.022	0.000		0.000	0.000	
000270KS	coeff 0.009	<b>15.571</b>	*** 20.9%	<b>0.040</b>	<b>69.534</b>	*** 19.9%	0.194	<b>2,908.307</b>	*** 19.3%
	T-stat 5.589	9.074		5.180	8.790		0.594	8.634	
	p-value 0.000	0.000		0.000	0.000		0.553	0.000	
000660KS	coeff 0.006	<b>30.586</b>	*** 30.2%	<b>0.022</b>	<b>145.943</b>	*** 32.7%	<b>-2.620</b>	<b>9,965.788</b>	*** 43.3%
	T-stat 3.505	11.408		2.605	12.078		-5.743	15.106	
	p-value 0.001	0.000		0.010	0.000		0.000	0.000	
009540KS	coeff 0.005	<b>11.879</b>	*** 27.1%	0.014	<b>55.872</b>	*** 30.0%	<b>-3.160</b>	<b>3,719.747</b>	*** 35.1%
	T-stat 2.391	10.762		1.614	11.546		-6.200	12.969	
	p-value 0.017	0.000		0.108	0.000		0.000	0.000	
015760KS	coeff 0.004	<b>13.329</b>	*** 30.0%	<b>0.021</b>	<b>53.100</b>	*** 27.5%	<b>-0.293</b>	<b>2,846.760</b>	*** 54.2%
	T-stat 3.786	11.555		4.722	10.881		-2.144	19.168	
	p-value 0.000	0.000		0.000	0.000		0.033	0.000	
017670KS	coeff 0.007	<b>5.602</b>	*** 23.1%	<b>0.030</b>	<b>22.857</b>	*** 23.4%	<b>-0.268</b>	<b>1,486.037</b>	*** 68.4%
	T-stat 8.848	9.672		9.450	9.758		-3.501	25.848	
	p-value 0.000	0.000		0.000	0.000		0.001	0.000	
030200KS	coeff 0.005	<b>10.287</b>	*** 32.1%	<b>0.023</b>	<b>43.154</b>	*** 31.8%	0.051	<b>2,266.557</b>	*** 65.5%
	T-stat 6.738	12.037		7.236	11.963		0.607	24.107	
	p-value 0.000	0.000		0.000	0.000		0.545	0.000	
033780KS	coeff 0.009	<b>3.941</b>	*** 10.0%	<b>0.037</b>	<b>18.741</b>	*** 12.7%	-0.193	<b>1,482.823</b>	*** 47.8%
	T-stat 9.794	5.964		9.100	6.787		-1.512	16.887	
	p-value 0.000	0.000		0.000	0.000		0.132	0.000	
051910KS	coeff 0.006	<b>11.581</b>	*** 24.7%	<b>0.020</b>	<b>52.941</b>	*** 27.0%	<b>-1.353</b>	<b>2,753.305</b>	*** 58.9%
	T-stat 2.966	9.962		2.392	10.576		-6.180	20.780	
	p-value 0.003	0.000		0.017	0.000		0.000	0.000	
066570KS	coeff 0.008	<b>13.151</b>	*** 26.4%	<b>0.040</b>	<b>53.817</b>	*** 22.4%	-0.339	<b>2,935.167</b>	*** 40.4%
	T-stat 6.123	10.500		6.387	9.430		-1.519	14.418	
	p-value 0.000	0.000		0.000	0.000		0.130	0.000	
086790KS	coeff 0.002	<b>17.851</b>	*** 35.5%	0.000	<b>83.987</b>	*** 38.4%	<b>-2.337</b>	<b>5,007.196</b>	*** 42.8%
	T-stat 1.029	13.049		0.039	13.866		-5.454	15.184	
	p-value 0.304	0.000		0.969	0.000		0.000	0.000	

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

**Table 5.2:** Linear Model: The Relationship between Volatility and Price Impact - continued

This table documents the coefficients of a simple linear model on the relationship between volatility and price impact (liquidity). As per hypothesis 2, we expect a positive and statistically significant relationship between the price impact of trades and stock market volatility. Here we examine three different proxies for volatility denoted  $\sigma_t^2$ : (a) the standard deviation of asset returns (b) the realized volatility using absolute daily returns and (c) the realized volatility using absolute high frequency transaction returns (tick returns). The measure of price impact  $\pi$  is as per our definition on eqn 24. The regression runs from Jan 2007 to Dec 2012 at a weekly frequency.

$$\sigma_t^2 = \beta_0 + \beta_1 \pi_t + \epsilon_t$$

Our summary shows that a linear positive relationship exists between volatility and price impact for all 25 Korean stocks in our sample. This is regardless of which volatility measure we use.

	(a) Standard deviation with daily returns			(b) Realized volatility with absolute daily returns			(c) Realized volatility with absolute tick returns		
	$\beta_0$	$\beta_1$	Adj. $R^2$	$\beta_0$	$\beta_1$	Adj. $R^2$	$\beta_0$	$\beta_1$	Adj. $R^2$
000720KS	coeff 1.075 T-stat 0.283 P-value	16.471 12.885 0.000	*** *** ***	0.005 0.599 0.549	74.267 13.251 0.000	*** *** ***	-1.147 -3.171 0.002	3,582.204 13.814 0.000	*** *** ***
000830KS	coeff 0.010 T-stat 9.629 P-value	4.198 4.198 0.000	*** *** ***	0.038 8.754 8.754	18.754 10.366 0.000	*** *** ***	-0.055 -0.430 0.667	1,111.867 20.722 0.000	*** *** ***
003550KS	coeff 0.006 T-stat 3.207 P-value	10.248 8.360 0.000	*** *** ***	0.025 2.877 0.004	46.217 8.586 0.000	*** *** ***	-1.094 -3.562 0.000	2,748.762 14.432 0.000	*** *** ***
009150KS	coeff 0.008 T-stat 5.810 P-value	12.238 11.267 0.000	*** *** ***	0.037 6.457 6.457	49.993 10.533 0.000	*** *** ***	-0.713 -2.937 15.053	2,999.022 15.053 0.000	*** *** ***
010140KS	coeff -0.002 T-stat -0.514 P-value	24.136 9.125 0.000	*** *** ***	0.000 (0.014) 0.304	112.078 9.210 0.000	*** *** ***	-3.519 -4.037 0.000	6,566.716 8.162 0.000	*** *** ***
010950KS	coeff 0.008 T-stat 4.803 P-value	6.911 6.911 0.000	*** *** ***	0.029 4.379 4.379	31.871 8.135 0.000	*** *** ***	-0.427 -3.538 0.000	1,609.639 22.577 0.000	*** *** ***
034220KS	coeff 0.007 T-stat 5.733 P-value	16.759 13.037 0.000	*** *** ***	0.033 5.669 0.000	72.355 12.457 0.000	*** *** ***	-0.902 -2.416 0.016	4,682.336 12.529 0.000	*** *** ***
035250KS	coeff 0.007 T-stat 4.404 P-value	6.971 6.971 0.000	*** *** ***	0.027 4.053 4.053	30.519 8.471 0.000	*** *** ***	-0.618 -1.718 0.087	1,605.001 8.340 0.000	*** *** ***
051900KS	coeff 0.009 T-stat 8.161 P-value	2.828 8.861 0.000	*** *** ***	0.036 6.196 6.196	11.641 8.969 0.000	*** *** ***	-0.759 -3.949 0.000	705.033 16.374 0.000	*** *** ***
Summary	coeff 0.006 % significant	12.383 100.0%	24.4%	0.024 76.0%	55.056 100.0%	24.9%	-1.014 76.0%	3,152.584 100.0%	43.6%

### 5.6 Conclusions

In conclusion, this chapter provides a simple theoretical model showing Kyle's  $\lambda$  is positively related to informational volatility. Since price impact as a proxy for Kyle's  $\lambda$  parameter, there should similarly be a positive relationship between it and asset price volatility. Our empirical test, using time-varying price impact prove this is indeed the case, and that the price impact of a trade is higher where informational variability is high.

## 5. THE PRICE IMPACT OF A TRADE AND ITS LINKAGE WITH VOLATILITY

---

# References

- [1] Amihud, Y. and H. Mendelson (1980) Dealership market: market-making with inventory, *Journal of Financial Economics* 8, 31 - 53
- [2] Andersen, T.G. (1996) Return volatility and trading volume: An information flow interpretation of stochastic volatility, *Journal of Finance* 51, 169 - 204
- [3] Andersen, T.G. and T. Bollerslev (1998) Answering the skeptics: Yes, standard volatility model do provide accurate forecasts, *Journal of International Economic Review* 39, 885 - 905
- [4] Baker, M. and Stein, J.C. (2004) Market liquidity as a sentiment indicator, *Journal of Financial Markets* 7, 271 - 299
- [5] Brownlees, C.T. and G.M. Gallo (2006) Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics & Data Analysis* 51, 2232 - 2245
- [6] Clark, B.K. (1973) A subordinated stochastic process model with finite variance for speculative prices, *Econometrica* 41, 135 - 155
- [7] Copeland, T.E. and D. Galai (1983) Information effects on the bid-ask spread, *Journal of Finance* 38, 1457 - 1469
- [8] de Jong, F., Nijman, T. and A. Roell (1995) A comparison of the cost of trading French shares on the Paris Bourse and on SEAQ International, *European Economic Review* 39, 1277 - 1301

## REFERENCES

---

- [9] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [10] Dufour, A. and R.F. Engle (2000a) The ACD model: predictability of the time between consecutive trades, *ISMA Centre, Discussion papers in Finance*, The University of Reading, UK
- [11] Ellis, K., Michaely, R. and M. O'Hara (2000) The accuracy of trade classification rules: evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* 35 - 4, 529 - 551
- [12] Foster, F.D. and S. Viswanathan (1993) Variations in trading volume, return volatility and trading costs: evidence on recent price formation models, *Journal of Finance* 48, 187 - 211
- [13] Grammig, J., Theissen, E. and O. Wünsche (2011) Time and the price impact of a trade: a structural approach, working paper, University of Tübingen, Germany
- [14] Grossman, S. J. (1976) On the efficiency of competitive stock markets where traders have diverse information, *Journal of Finance* 31, 573 - 585
- [15] Grossman, S. J. and J. E. Stiglitz(1980) On the impossibility of informationally efficient markets, *The American Economic Review* 70 - 3, 393-408
- [16] Glosten, L. and L. H. Harris (1988) Estimating the components of the bid-ask spread, *Journal of Financial Economics* 21, 123 - 142
- [17] Glosten, L. and P. Milgrom (1985) Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 14, 71 - 100
- [18] Harrison, M. and D. Kreps (1978) Speculative investor behaviour in a stock market with heterogeneous expectations, *Quarterly Journal of Economics* 92, 323 -336
- [19] Hasbrouck, J. (1991) Measuring the information content of stock trades, *Journal of Finance* 46, 179 - 208

## REFERENCES

---

- [20] Hasbrouck, J. (1991a) The summary informativeness of stock trades: an econometric analysis, *Review of Financial Studies* 4 No.3, 571 - 595
- [21] Hong, H., Scheinkman, J. and W. Xiong (2006) Asset float and speculative bubbles, *Journal of Finance* 61, 1073 - 1117
- [22] Huang, R.D. and H.R. Stoll (1997) The components of the bid-ask spread: a general approach, *Review of Financial Studies* 10, 995 - 1034
- [23] Ho, T. and H.R. Stoll (1980) On dealer markets under competition, *Journal of Finance* 35, 259 - 267
- [24] Ho, T. and H.R. Stoll (1981) Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 47 - 73
- [25] Ho, T. and H.R. Stoll (1983) The dynamics of dealer markets under competition, *Journal of Finance* 38, 1053 - 1074
- [26] Jang, H. and P.C. Venkatesh (1991) Consistency between predicted and actual bid-ask quote revisions, *Journal of Finance* 46, 433 - 446
- [27] Kraemer, N. and J. Schaefer (2011) Package 'parcor', *CRAN repository*
- [28] Kyle, A.S. (1985) Continuous auctions and insider trading, *Econometrica* 53 - 6, 1315 - 1336
- [29] Lee, C. and M. Ready (1991) Inferring trade direction from intraday data, *Journal of Finance* 46, 733 - 746
- [30] Escribano, A., Pascual, R. and M. Tapia (2004) Adverse selection costs, trading activity and price discovery in the NYSE: An empirical analysis *Journal of Banking & Finance* 28, 107 - 128
- [31] Stoll, H.R. (1978) The supply of dealer services in securities markets, *Journal of Finance* 33, 1133 - 1151



## REFERENCES

---

- [32] Van Ness, B.F., Van Ness, R.A. and R.S. Warr (2002) Is the adverse selection component really higher on the NYSE/Amex than on the Nasdaq? *Journal of Business Finance and Accounting* 29 807 - 824

## 6

# The Price Impact of a Trade, VPIN and the Role of Informed Trader Heterogeneity

1

**Abstract:** *We document a curious result: both price impact and volume synchronized probability of informed trading (VPIN) relate to information asymmetry and adverse selection, yet empirically we find a statistically significant negative correlation between the two. The price impact of a trade measures the information content of incoming order flow and the risk aversion costs of being picked-off by an informed trader. Price impact captures the adjustment made by market makers and liquidity traders upon observing active order flow. VPIN measures order flow imbalance and the probability of being adversely selected. We conduct VAR modeling and show a contemporaneous negative relationship between price impact and VPIN, but no significant lag or lead. We provide a theoretical explanation showing that in cases where there is heterogeneous beliefs amongst informed or active traders, then a high price impact can lower order imbalance and result in a lower VPIN. Therefore, we show price impact and VPIN behaves differently despite having similar objectives on measuring information asymmetry and adverse selection.*

---

<sup>1</sup>Chapter 6 and its extensions form part of a working paper with Dr. Quan Gan titled "The Tale of Two Measures: Price Impact and VPIN"

### 6.1 Introduction

Chung, Li and McNish (2005) found it surprising that little direct evidence exist on the relationship between information-based trading and price impact. Using Easley, Kiefer and O'Hara's (1997) probability of informed trading (PIN) and Dufour and Engle's (2000) price impact measure, they show that a significant positive relationship exists on a cross-sectional basis using 538 NYSE listed stocks. We ask whether a similar relationship holds from a time-series perspective.

Whilst PIN has been used in numerous studies cross-sectionally (e.g., Chen and Zhao, 2012), few have considered it from a time-series perspective. This rests on the estimation methodology of PIN, which uses static parameters. Easley, Engle, O'Hara and Wu (2008) and Tay, Ting, Tse and Warachka (2009) suggest modifying PIN by incorporating GARCH-like features for its parameters. However recently, Easley, Lopez de Prado and O'Hara (2011, 2012) utilized a heuristic discovered in Easley, Engle, O'Hara and Wu (2008) to develop volume synchronized probability of informed trading (VPIN) which uses high frequency rather than daily data to obtain a time-varying measure for order flow toxicity. This methodology also does not entail maximum likelihood estimation and bypasses the issues, such as floating point exception, raised by Lin and Ke (2011). The development of VPIN provides us with an opportunity to test its relationship with price impact in time-series.

We examine the relationship between these two measures, price impact (Hasbrouck, 1991; Dufour and Engle, 2000; Escibano, Pascual and Tapia, 2004) and VPIN ( Easley, Lopez de Prado and O'Hara 2011, 2012) using high frequency tick data. The price impact of a trade is often used to measure adverse selection costs (see Escibano, Pascual and Tapia, 2004) and VPIN, a high frequency version of PIN, measures the likelihood of liquidity providers being adversely selected. In a Goldman Sachs report, Jeria and Sofianos (2008) defines adverse selection to be the natural tendency for passive orders (in the limit order book) to fill quickly when they should fill slowly and fill slowly when they should fill quickly. In essence, VPIN examines the magnitude of order imbalance in volume time. Higher levels of order imbalance would be an

indication of greater adverse selection. Easley, Lopez de Prado and O'Hara (2011) state that in periods where there is a lot of information-based trades, VPIN will be large. Furthermore, they show VPIN "sets the stage" for illiquidity through the example of flash crashes.

Hasbrouck (1991) state high price impact is a sign of increased information content and greater levels of information-based trading. Naturally the concept of information content and adverse selection are related. So much so that Escribano, Pascual and Tapia (2004) use price impact, a measure Hasbrouck (1991) uses for information content, to measure adverse selection costs. Intuitively, higher information content will result in adverse selection and illiquidity. If the market makers or liquidity traders are aware of a greater inflow of informed trading, they are likely to increase the magnitude of quote revisions to account for the greater adverse selection. This idea is mentioned as far back as Kyle (1985). Since VPIN and price impact are both measuring asymmetric information, adverse selection and illiquidity, it is reasonable to expect that they should be positively correlated from a time-series perspective.

However, this is in fact not the case. We provide empirical evidence and a theoretical explanation, involving heterogeneity amongst informed traders, to support this bizarre finding.

Firstly, we document the methodology used to estimate price impact and VPIN. Our method for estimating price impact extends upon Dufour and Engle's (2000) VAR model to account for durations, spreads and depth (in chapter 4 we show that order book illiquidity can influence price impact). In a study by Chakrabarty, Pascual and Shkilko (2013) it is shown that tick based rules to be superior to bulk volume classification (BVC) using data from NASDAQ's INET platform. Therefore, we make an adjustment in Easley, Lopez de Prado and O'Hara's (2012) VPIN methodology to also use Ellis, Michaely and O'Hara's (2000) trade classification algorithm as opposed to BVC. This also creates consistency when comparing it with price impact (which also uses Ellis, Michaely and O'Hara's (2000) method).

We conduct time-series analysis between price impact and VPIN on a selection of 24 large

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

capitalization Korean stocks from January 2007 to December 2012. We find a statistically significant negative correlation between price impact and VPIN. Intuitively we suggest that order flow imbalance or toxicity cannot materialize in periods where price impact is high. To confirm our model predictions, we conduct a robust sparse VAR model on price impact, VPIN, volume and volatility. We find that there is a strong contemporaneous relationship between price impact and VPIN. Whilst price impact is positively related to contemporaneous volatility (as documented earlier in chapter 5), VPIN was neither driven by volatility nor volume.

This chapter is organized as follows. Section 6.2 describes the dataset we use. Section 6.3 provides details on VPIN estimation. Section 6.4 documents our methodology for price impact utilizing adaptive lasso regularization. Section 6.5 documents the negative Pearson's correlation between price impact and VPIN. Section 6.6 provides the empirical results using a sparse VAR model. Section 6.7 provides our theoretical explanation for this phenomenon and section 6.8 concludes.

### 6.2 Data

The dataset we use for this research comes from Thomson Reuters Tick History and records trade and quote data from the Korea Exchange to the microsecond. Our tick history dataset begins from January 2007 and ends at December 2012. We pick 24 of the largest capitalization stocks on the Korea Exchange with complete tick data history, this is shown in chapter 2. High frequency data anomalies are cleaned using the Brownlees and Gallo (2006) approach <sup>1</sup>.

Trade initiation identifies whether the trade was initiated by the buyer or by the seller<sup>2</sup>. For

---

<sup>1</sup>This is similar to (but not the same as) using high frequency Bollinger bands of 2 standard deviations. For each transaction price, a window of 41 transactions is constructed: 20 prior transactions and 20 post transactions. If the difference between the transaction price and the average price of the window is greater than 2 standard deviations, it is discarded. For precise details see Brownlees and Gallo (2006)

<sup>2</sup>Tick data from most North American and Asian exchanges do not have qualifier tags that determine trade initialization. Therefore several algorithms have been used by researchers to tackle this issue

---

### 6.3 Volume Synchronized Probability of Informed Trading

our research, we have used the Ellis, Michaely and O'Hara (2000) approach <sup>1</sup>. To the best of our knowledge this has been the most recent trade initialization algorithm which correctly classifies are larger percentage of trades than the popular Lee and Ready (1991) algorithm or the tick approach.

### 6.3 Volume Synchronized Probability of Informed Trading

VPIN was recently developed by Easley, Lopez de Prado and O'Hara (2011, 2012) specifically for high frequency data. It relates to a long series of existing literature on the topic of the probability of informed trading (see Easley, Kiefer, O'Hara and Paperman ,1996; Easley, Kiefer and O'Hara, 1997; Easley, Hvidkjaer and O'Hara, 2002). In Easley, Kiefer, O'Hara and Paperman (1996), the PIN model (henceforth EKOP PIN) using daily aggregate buy and sell imbalance was developed. EKOP PIN assumes news events are drawn on a daily basis from a Bernoulli random variable  $Bin(1,p)$ , with a constant parameter  $p$ . Likewise the conditional event of bad news for a particular day is drawn from another constant parameter Bernoulli random variable. Buy and sell initiated trades are assumed to be transacted with exponential waiting time; therefore for a given time interval the aggregate number of buys or sells can be modeled using Poisson distributions. Two Poisson processes are required, one for buys and one for sells. The intensity of each process is determined by the presence of informed and uninformed traders. For instance, informed traders will only be buying if there is good news and selling if there is bad news. In essence, the EKOP PIN model is a Bernoulli modulated Poisson process. The PIN metric itself in EKOP PIN relates to the expected percentage of informed trading. It is well known to be,

$$PIN = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon} \quad (6.1)$$

where  $\alpha$  is the probability of a news event,  $\mu$  is the intensity of informed trading and  $\varepsilon$  is the intensity of the uninformed traders.

---

<sup>1</sup>All trades executed at the ask quote are classified as a buy initiation. All trades executed at the bid quote are classified as a sell initiation. All other trades are categorized by the tick rule

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

In Easley, Engle, O'Hara and Wu (2008), a time-varying PIN model was developed. This model uses time-varying intensity rates and probabilities (rather than static intensities and probabilities in EKOP PIN). They show that for a particular interval frequency (e.g. days),  $\alpha\mu$  can be approximated as  $E(V_t^{sell} - V_t^{buy})$  where  $V_t^{sell}$  and  $V_t^{buy}$  are the aggregate sell initiated volume and buy initiated volume in interval  $t$ . Likewise the denominator for PIN, i.e.,  $\alpha\mu + 2\varepsilon$ , can be approximated as  $E(V_t^{sell} + V_t^{buy})$  (see Easley, Lopez de Prado and O'Hara, 2012). VPIN draws upon these two approximations, creating what can loosely be described as a high frequency PIN estimate. As discussed in Easley, Lopez de Prado and O'Hara (2012), the VPIN merit over other PIN approaches is that it completely bypasses the maximum likelihood procedure and any difficulties associated with it, such as the floating point exception (see Lin and Ke, 2011).

Instead of clock-time intervals in EKOP PIN, VPIN uses volume-time. The application of subordinated stochastic processes in finance was explored initially by Clark (1973) with a volume subordinator and subsequently Zhou (1996) with volatility subordinators and Ane and Geman (2000) with number of trades. The argument involves sampling the time-series at a frequency or speed that better matches the speed of information arrival. Easley et al. (2012) show by using volume buckets (i.e. volume-time), sample volatility clustering is reduced. Easley, Lopez de Prado and O'Hara (2012) methodology involves absolute order flow imbalance. The measure is defined as,

$$VPIN = \frac{\sum_{t=1}^n |V_t^{sell} - V_t^{buy}|}{nV} \quad (6.2)$$

where  $V$  is the total volume in each bucket (such that  $V = V_t^{sell} + V_t^{buy}$ ) and  $t$  denotes the buckets. In Easley et al. (2012),  $n = 50$ , such that VPIN is estimated with 50 buckets. Each volume bucket is  $\frac{1}{50}$  of average daily total volume. Therefore, it corresponds more or less to finding daily VPIN. It can be seen that the Easley, Lopez de Prado and O'Hara (2012) methodology is essentially a ratio of order flow imbalance over total volume.

In Easley et al. (2012), BVR is used to determine buy and sell initiated trades. BVR provides greater simplicity over the traditional tick-based approaches used by market microstructure

### 6.3 Volume Synchronized Probability of Informed Trading

---

practitioners (such as the tick rule or the Lee and Ready, 1991, algorithm). The process does not require trade by trade tick data, instead only 1-minute frequency price / volume series are required (i.e. time bars and volume bars). Using the price change in each period  $\Delta P$ , a probabilistic percentage of buys and sells is determined, i.e. the percentage of buys is  $Z(\frac{\Delta P}{\sigma_{\Delta P}})$  and the percentage of sells is  $1 - Z(\frac{\Delta P}{\sigma_{\Delta P}})$ . Whilst it is clear that BVR is computationally less intensive, Chakrabarty, Pascual and Shkilko (2013) find that it is inferior in terms of accuracy when compared to tick based rules. Using Nasdaq's INET order book, they show that the basic tick rule is more accurate than BVR. From their dataset, they find BVR is most accurate with time bars of 1 hour frequency; and under such settings BVC classifies 79.7% of volume correctly compared to 90.8% for the tick rule. Therefore, in light of Chakrabarty, Pascual and Shkilko (2013), we decide to modify the Easley et al. (2012) method and classify buy and sell initiation via a tick-based method. To be consistent with the methodology we employed later for price impact, we use the Ellis, Michaely and O'Hara (2000) method, which is superior to the tick rule or Lee and Ready (1991) method. The dataset we use is identical to the one we use for price impact.

In a stationary scenario with information homogeneity across time, it may be concluded that having larger volume buckets produce greater precision for VPIN estimation. However, in reality information flow is time dependent and therefore by utilizing larger volume buckets we reduce the accuracy and granularity of VPIN to reflect underlying information. Therefore, bucket size determination is tricky involving a trade-off: too small means VPIN is unlikely to be accurate due to discrete buy/sell volume realizations (upwards bias) and too large means you might lose possible information. Knowing this we find it extremely problematic to compare VPIN measures between stocks, since bucket size is determined at a per stock basis. Therefore in this paper, we do not compare VPIN cross-sectionally. Time series analysis on a single stock is not a problem since the bucket size  $V$  is consistent across time. We estimate VPIN for all stocks in our sample. We use bucket size corresponding to  $\frac{1}{25}$  of average daily total volume, as Korean equities are less frequently traded than E-mini S&P futures used in the Easley et al. (2012) paper. Below we provide an illustration of VPIN time-series for Samsung Electronics.

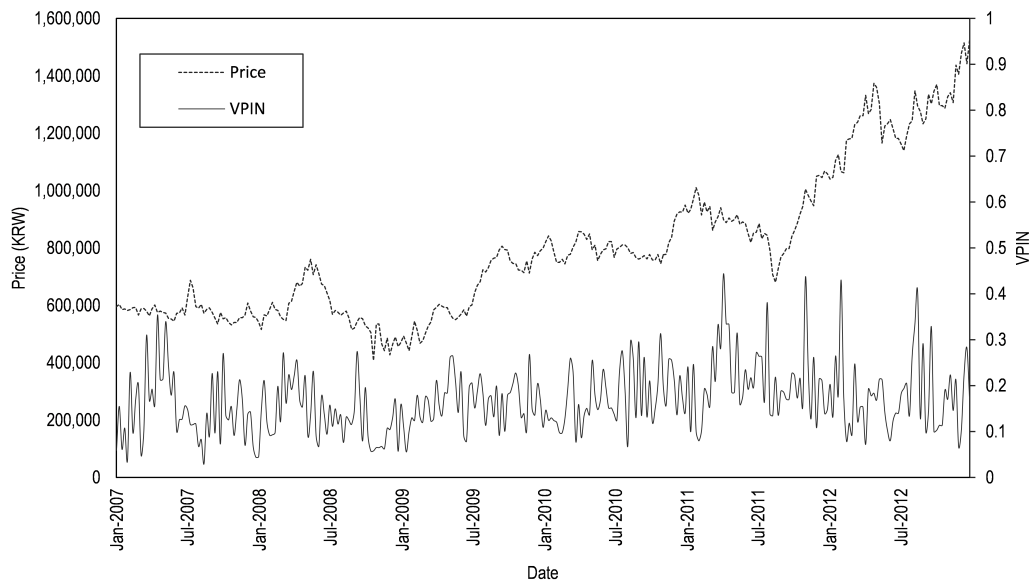


## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

**Figure 6.1:** Time-series VPIN: Samsung Electronics

VPIN and price series for Samsung Electronics from Jan-2007 to Dec-2012.



### 6.4 Price Impact of a Trade

As shown in the existing literature (see Hasbrouck, 1991; Dufour and Engle, 2000; Chung, Li, McNish, 2005), trades trigger quote revisions, and furthermore there is a lagged effect. Price impact is measured as the cumulative impulse response of quote revisions after a trade shock. In Hasbrouck (1991) a VAR model was considered with two endogenous variables: trades (signed volume) and quote revisions. Using this approach Hasbrouck (1991) shows that quote revisions have a lagged response to trades and that trades are significantly serially correlated. By generating cumulative impulse response functions from the VAR model one is able to analyze the magnitude of private information impounded into the price via trades, and therefore determine the information content of a trade. Dufour and Engle (2000) and Escibano, Pascual and Tapia (2004) provide extensions to the VAR model. Dufour and Engle (2000) incorporate an exogenous factor, durations, into the model. This motivation is driven from Diamond and Verrecchia (1987) and Easley and O'Hara (1992) who suggest longer durations represent no news. Informed traders will only enter the market if new information exists from which they can profit

against uninformed or liquidity traders. Both papers show longer durations have a negative impact on quote revisions, suggesting in periods where trading is fast, the price impact is also higher. In this study, we introduce a VARX (a VAR with exogenous variables) model which extends upon Dufour and Engle’s (2000) VAR model, considering not only durations, but also volume, bid-ask spreads and depth as well. We employ Ellis, Michaely and O’Hara (2000) trade classification algorithm which is shown to be more accurate than the commonly used Lee and Ready (1991) approach for determining trades initialization. Our model below extends upon the framework developed by Hasbrouck (1991) and Dufour and Engle (2000). As is discussed in Escribano, Pascual and Tapia (2004), our reduced-form approach accounts for the dynamic impact of trades and is superior to traditional structural models such as the Glosten model. We also introduce regularization, from machine learning, to market microstructure. Regularization imposes Occam’s razor to our VARX model and prevents over-fitting. In our study, we utilize a version of Tibshirani’s (1996) lasso (adaptive lasso - Zou, 2006) on VARX; this allows us to conduct parameter subset selection and estimation in a single operation. It is useful to note that Zou (2006) show adaptive lasso holds *oracle properties* - the estimated subset converges in probability to the true subset. As high frequency VAR models are generally quite verbose, we believe our cross-validated regularization approach provides a good solution in determining whether coefficients are truly significant.

Our VARX model is described as follows.

Two endogenous variables are considered,  $r_t$  and  $\dot{v}_t$ . Quote revisions  $r_t$  are defined to be the movement or changes in the midpoint price determined through the best bid and ask price in the order book  $r_t = 100 \times (\log(q_{t+1}) - \log(q_t))$  where  $q_t = \frac{q_t^{bid} + q_t^{ask}}{2}$ . Instead of transaction price, the use of the midpoint price  $q_t$  eliminates the bid-ask bounce associated with using returns generated through traded prices. The scaling factor of 100 is consistent with Dufour and Engle (2000).

Trades  $\dot{v}_t = x_t * \log v_t$  is signed trade volume, i.e. log volume multiplied by trade initialization. The log transformation is applied to reduce the impact of extraordinarily large volumes in

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

the dataset, this is suggested by Potters and Bouchaud (2003) and Hafner (2005) whilst studying statistical properties of the market impact of trades.

Several exogenous variables are considered. Firstly, we consider the impact of trade durations  $d_t$ , as it is shown in the literature (Diamond and Verrecchia, 1987, and Easley and O'Hara, 1992) that durations (trading intensity) have a negative (positive) effect price impact. The rationale is that trading intensity increases in periods of greater information, and therefore each trade would contain greater information content, and subsequently greater price impact. This was confirmed in Dufour and Engle (2000).

Secondly, we consider the bid-ask spread right before the trade,  $s_t$ , which is a common measure of static liquidity. And thirdly, we consider the first level depth of the orderbook right before the trade  $h_t|x_t$ . This second measure of liquidity is conditional on trade direction  $x_t$ . For a buy trade, we consider the first level depth of the ask side, and for a sell we consider the first level depth of the bid side. The inclusion of these measures is because we wish to examine the effect of market liquidity on price and size impact.

Pascual, Escribano and Tapia (2004) show a generalization of Hasbrouck (1991). Our model draws upon the same framework.

$$\begin{aligned} r_t &= \sum_{i=1}^{\infty} \gamma_i^{(1,1)} r_{t-i} + \sum_{i=0}^{\infty} \left[ \gamma_i^{(1,2)} + \beta_i^{(1,1)} d_{t-i} + \beta_i^{(1,2)} s_{t-i} + \beta_i^{(1,3)} h_{t-i}|x_{t-i} \right] \dot{v}_{t-i} + \varepsilon_t^{(1)} \\ \dot{v}_t &= \sum_{i=1}^{\infty} \gamma_i^{(2,1)} r_{t-i} + \sum_{i=1}^{\infty} \left[ \gamma_i^{(2,2)} + \beta_i^{(2,1)} d_{t-i} + \beta_i^{(2,2)} s_{t-i} + \beta_i^{(2,3)} h_{t-i}|x_{t-i} \right] \dot{v}_{t-i} + \varepsilon_t^{(2)} \end{aligned} \quad (6.3)$$

It can be interpreted that  $d_{t-i}$ ,  $s_{t-i}$  and  $h_{t-i}|x_{t-i}$  are control variables, that will impact the relationship between trades and quote revisions. This can be re-expressed in VARX format with  $p$  lags.

$$\begin{aligned}
 \begin{bmatrix} 1 & -\gamma_0^{(1,2)} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} r_t \\ \dot{v}_t \end{bmatrix} &= \begin{bmatrix} \gamma_1^{(1,1)} & \gamma_1^{(1,2)} \\ \gamma_1^{(2,1)} & \gamma_1^{(2,2)} \end{bmatrix} \begin{bmatrix} r_{t-1} \\ \dot{v}_{t-1} \end{bmatrix} + \dots + \begin{bmatrix} \gamma_p^{(1,1)} & \gamma_p^{(1,2)} \\ \gamma_p^{(2,1)} & \gamma_p^{(2,2)} \end{bmatrix} \begin{bmatrix} r_{t-p} \\ \dot{v}_{t-p} \end{bmatrix} \\
 &+ \begin{bmatrix} \beta_0^{(1,1)} \\ 0 \end{bmatrix} d_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,1)} \\ \beta_1^{(2,1)} \end{bmatrix} d_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,1)} \\ \beta_p^{(2,1)} \end{bmatrix} d_{t-p} \dot{v}_{t-p} \\
 &+ \begin{bmatrix} \beta_0^{(1,2)} \\ 0 \end{bmatrix} s_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,2)} \\ \beta_1^{(2,2)} \end{bmatrix} s_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,2)} \\ \beta_p^{(2,2)} \end{bmatrix} s_{t-p} \dot{v}_{t-p} \\
 &+ \begin{bmatrix} \beta_0^{(1,3)} \\ 0 \end{bmatrix} h_t |x_t \dot{v}_t + \begin{bmatrix} \beta_1^{(1,3)} \\ \beta_1^{(2,3)} \end{bmatrix} h_{t-1} |x_{t-1} \dot{v}_{t-1} + \dots + \begin{bmatrix} \beta_p^{(1,3)} \\ \beta_p^{(2,3)} \end{bmatrix} h_{t-p} |x_{t-p} \dot{v}_{t-p} \\
 &+ \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}
 \end{aligned} \tag{6.4}$$

Let  $\mathbf{Y}_t = (r_t, \dot{v}_t)^T$ . Our VAR model is re-expressed as,

$$\begin{aligned}
 \mathbf{Y}_t &= \mathbf{A}_0 \mathbf{Y}_t + \dots + \mathbf{A}_p \mathbf{Y}_{t-p} + \mathbf{B}_0 d_t \dot{v}_t + \dots + \mathbf{B}_p d_{t-p} \dot{v}_{t-p} \\
 &+ \mathbf{C}_0 s_t \dot{v}_t + \dots + \mathbf{C}_p s_{t-p} \dot{v}_{t-p} + \mathbf{D}_0 (h_t |x_t) \dot{v}_t + \dots + \mathbf{D}_p (h_{t-p} |x_{t-p}) \dot{v}_{t-p} + \boldsymbol{\varepsilon}_t
 \end{aligned}$$

where,

$$\mathbf{A}_0 = \begin{bmatrix} 0 & \gamma_0^{(1,2)} \\ 0 & 0 \end{bmatrix} \quad \mathbf{B}_0 = \begin{bmatrix} \beta_0^{(1,1)} \\ 0 \end{bmatrix} \quad \mathbf{C}_0 = \begin{bmatrix} \beta_0^{(1,2)} \\ 0 \end{bmatrix} \quad \mathbf{D}_0 = \begin{bmatrix} \beta_0^{(1,3)} \\ 0 \end{bmatrix}$$

and for  $k = 1 \dots p$ ,

$$\mathbf{A}_k = \begin{bmatrix} \gamma_k^{(1,1)} & \gamma_k^{(1,2)} \\ \gamma_k^{(2,1)} & \gamma_k^{(2,2)} \end{bmatrix} \quad \mathbf{B}_k = \begin{bmatrix} \beta_k^{(1,1)} \\ \beta_k^{(2,1)} \end{bmatrix} \quad \mathbf{C}_k = \begin{bmatrix} \beta_k^{(1,2)} \\ \beta_k^{(2,2)} \end{bmatrix} \quad \mathbf{D}_k = \begin{bmatrix} \beta_k^{(1,3)} \\ \beta_k^{(2,3)} \end{bmatrix}$$

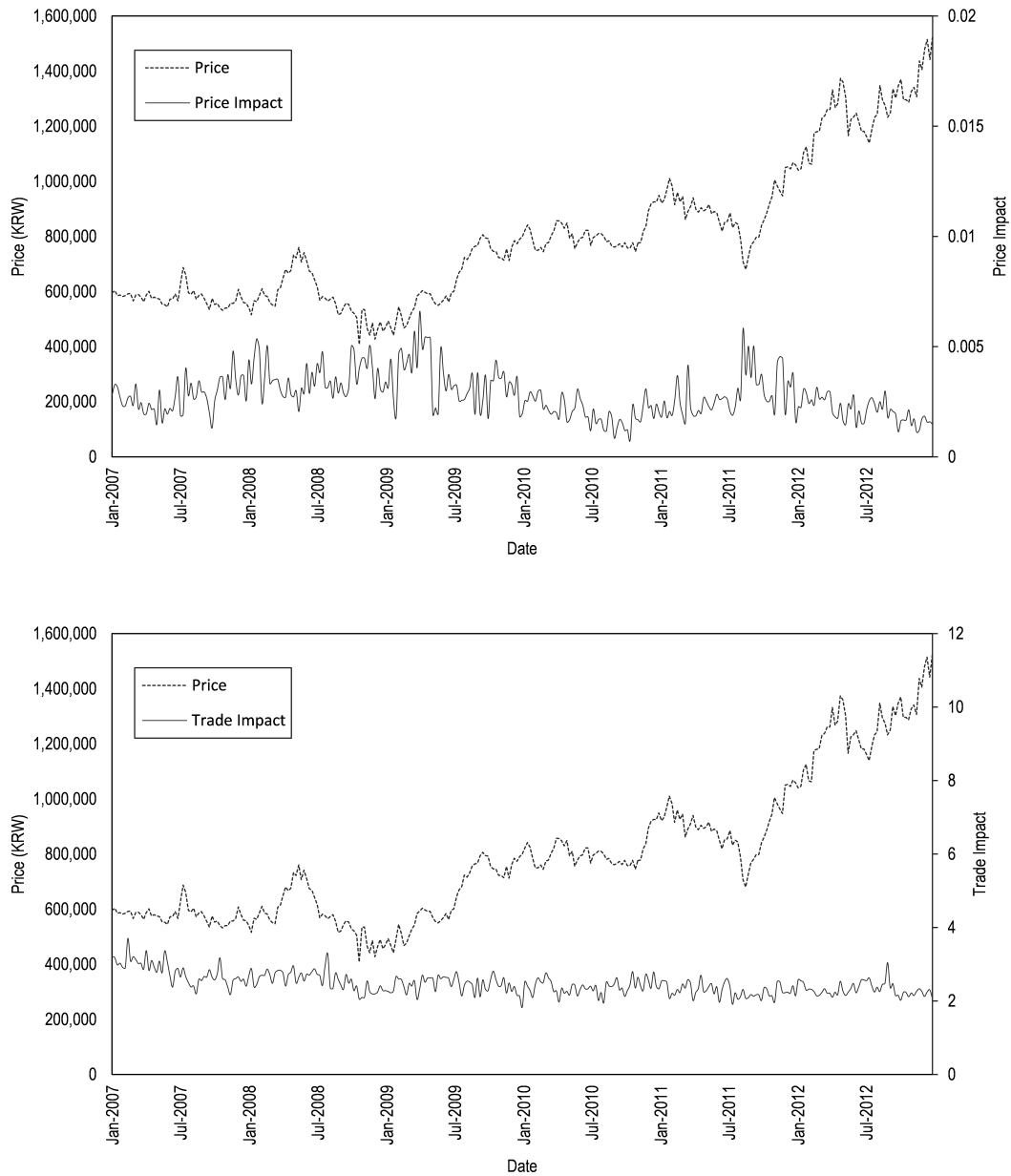
and  $\boldsymbol{\varepsilon}_t$  is a white noise process with the covariance matrix  $E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^T) = \Sigma_\varepsilon$ .

Further details on the estimation of price impact has been covered in chapter 4. Below we plot price impact (and also trade impact), derived from the CIRF of our VAR model across time, at weekly frequency for Samsung Electronics.

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

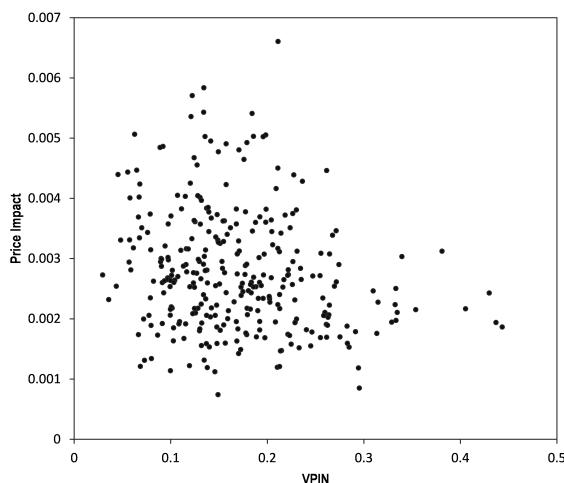
**Figure 6.2:** Time-series Price Impact and Trade Impact

Price impact, trade impact and price series for Samsung Electronics from Jan-2007 to Dec-2012.



**Figure 6.3:** Scatter-plot of Price Impact and VPIN

Samsung Electronics scatter-plot of weekly frequency price impact and VPIN for the period January 2007 to December 2012.



## 6.5 The Curious Case of Negative Correlation

Both price impact and VPIN claim to measure adverse selection (see Escribano, Pascual, and Tapia, 2004, for price impact and Easley, Lopez de Prado and O'Hara 2011, 2012, for VPIN). Price impact is concerned with measuring the magnitude of quote revisions as an indication of adverse selection, whilst VPIN is concerned with level of order imbalance. We estimate both price impact and VPIN on a weekly frequency. The former involves using weekly blocks of high frequency tick data to estimate the parameters of the adaptive lasso VARX model. The latter involves taking the average VPIN estimate over the period of a week.

We find that they are negatively correlated. If both measures were precisely measuring adverse selection, then one would expect a high degree of positive correlation between price impact and VPIN. The fact that this is not the case suggests that they are in fact unrelated measures, and at least one of them is not measuring adverse selection, or that there consists of more dimensions or stages to adverse selection. For a typical Korean large capitalization stock, Samsung Electronics, the Pearson's correlation between price impact and VPIN is -0.197, and this is significant at the 1% level. On average for all stocks, we find the correlation to be -0.313. The average p-value is 1.7%.

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

In this initial probe, we estimate their correlations between each other and with other trading variables. Firstly, we consider *trade impact*. This is the mirror of price impact as it is calculated from the cumulative impulse response of trades from a trade, whereas price impact is the cumulative impulse response of quote revisions from a trade. Trade impact detects trading momentum or trade autocorrelation, i.e., consecutive buys or sells in sequence. Secondly, we consider volume and trade size. Volume refers to the total traded volume in the weekly block. Trade size refers to the average transaction size. Thirdly, we consider a variety of volatility measures. In Table 5, Daily  $\sigma$  refers to the standard deviation of daily price, whilst tick  $\sigma$  refers to the standard deviation of all quote revisions in the weekly block. Daily RV denotes the realized volatility of daily price and tick RV denotes the realized volatility of quote revisions.

Table 6.1 provides correlations for a single stock, Samsung Electronics, and table 6.2 provides average correlations across all 24 stocks in our sample. We find that price impact is positively correlated with volume, trade size and a selection of volatility measures. On the other hand, we find that VPIN is negatively correlated to volume and trade size, and negatively correlated with three of the four volatility measures. Also, trade impact is an uninteresting variable, not related with either price impact or VPIN.

**Table 6.1:** Pearson Correlations for Samsung Electronics

Here we provide the Pearson correlations and its corresponding p-values for a selection of metrics on Samsung Electronics (005930 KS). The correlation is based on weekly measurements from Jan 2007 to Dec 2012. Price impact and trade impact are derived from adaptive lasso VARX cumulative impulse response functions. VPIN is estimated with  $n = 25$  (i.e. an average of 25 buckets per day). Four measures for volatility is provided. Two standard deviation measures, one on daily frequency and one using quote revisions. RV is short for realized volatility, which is calculated as the sum of absolute returns.

	Price Impact	VPIN	Trade Impact	Volume	Trade Size	Daily $\sigma$	Volatility Tick $\sigma$	Daily RV	Tick RV
Price Impact	1	<b>-0.197</b> <b>0.000</b>	<b>0.111</b> <b>0.050</b>	<b>0.422</b> <b>0.000</b>	<b>0.349</b> <b>0.000</b>	<b>0.400</b> <b>0.000</b>	<b>0.315</b> <b>0.000</b>	<b>0.415</b> <b>0.000</b>	<b>0.599</b> <b>0.000</b>
VPIN	1		-0.013 0.815	<b>-0.120</b> <b>0.033</b>	<b>-0.303</b> <b>0.000</b>	<b>-0.192</b> <b>0.001</b>	-0.071 0.211	<b>-0.160</b> <b>0.005</b>	<b>-0.202</b> <b>0.000</b>
Trade Impact			1	<b>0.337</b> <b>0.000</b>	<b>0.594</b> <b>0.000</b>	-0.010 0.855	0.065 0.248	-0.029 0.603	<b>-0.145</b> <b>0.010</b>
Volume				1	<b>0.605</b> <b>0.000</b>	<b>0.529</b> <b>0.000</b>	<b>0.225</b> <b>0.000</b>	<b>0.637</b> <b>0.000</b>	<b>0.655</b> <b>0.000</b>
Trade Size					1	<b>0.146</b> <b>0.009</b>	<b>0.239</b> <b>0.000</b>	<b>0.167</b> <b>0.003</b>	<b>0.180</b> <b>0.001</b>
Daily $\sigma$						1	<b>0.302</b> <b>0.000</b>	<b>0.916</b> <b>0.000</b>	<b>0.633</b> <b>0.000</b>
Tick $\sigma$							1	<b>0.287</b> <b>0.000</b>	<b>0.349</b> <b>0.000</b>
Daily RV								1	<b>0.735</b> <b>0.000</b>
Tick RV									1



## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

Our findings suggest that high price impact and large quote revisions are not associated with high buy-sell order imbalance. We find price impact is higher in periods with greater trading volume and greater volatility. This seems to suggest that price impact is related to information flow. The intensity of underlying information flow is commonly associated with volume or volatility. For example, when choosing subordinators for information flow synchronization, Clark (1973) uses volume, Zou (1996) uses volatility and Ane and Geman (2000) uses number of trades. The positive association between price impact and information flow supports Hasbrouck's (1991) notion where price impact is viewed as a proxy for the level of information content. Most certainly, if there is greater underlying information, the information impounding process would result in greater volume, volatility, trade size and price impact. On the other hand, we note VPIN is negatively correlated to volume and volatility. At first, this seems counterintuitive, as it seems to suggest that VPIN is higher in periods with less information flow and lower vice versa. We find that VPIN is higher in periods with smaller trade sizes, so it may be the case that large buy orders are being broken into smaller consecutive trades. However, this is not the case as consecutive trading in the same direction would lead to a greater trade impact; VPIN is not correlated with trade impact. Whilst it is hard to believe that high VPIN and high order imbalance is associated with periods of low price impact, volume and volatility. It perhaps makes more sense to think in reverse: would order imbalance materialize in periods of high price impact, volume and volatility? We argue that the very reason order imbalance has materialized is the fact that there exists a reasonable liquid market with low price impact. If price impact is high, then a single buy trade is likely to trigger quotes to be raised high, this deters further buy trading and encourages sell initiated trades.

**Table 6.2:** Average Pearson Correlations for All Stocks

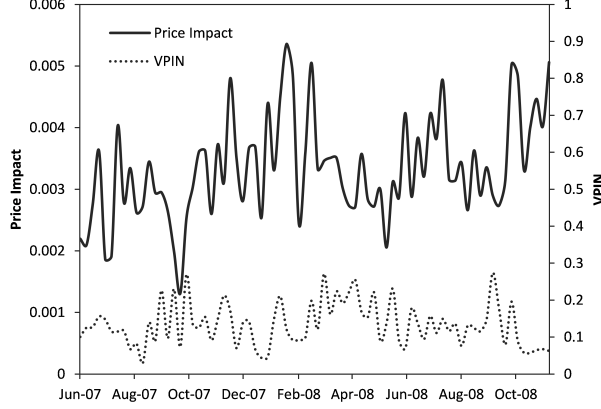
Here we provide the average Pearson correlations and its corresponding p-values for the all the stocks in our sample. The correlation is based on weekly measurements from Jan 2007 to Dec 2012. Price impact and trade impact are derived from adaptive lasso VARX cumulative impulse response functions. VPIN is estimated with  $n = 25$  (i.e. an average of 25 buckets per day). Four measures for volatility is provided. Two standard deviation measures, one on daily frequency and one using quote revisions. RV is short for realized volatility, which is calculated as the sum of absolute returns. A significant

	Price Impact	VPIN	Trade Impact	Volume	Trade Size	Daily $\sigma$	Tick $\sigma$	Volatility Daily RV	Tick RV
Price Impact	1	<b>-0.313</b> <b>0.017</b>	0.151 0.131	0.274 0.089	0.254 0.077	<b>0.484</b> <b>0.000</b>	<b>0.335</b> <b>0.010</b>	<b>0.490</b> <b>0.000</b>	<b>0.599</b> <b>0.000</b>
VPIN		1	0.162 0.149	-0.052 0.216	<b>-0.349</b> <b>0.017</b>	-0.153 0.066	<b>-0.172</b> <b>0.042</b>	-0.142 0.073	<b>-0.205</b> <b>0.030</b>
Trade Impact			1	0.260 0.052	0.045 0.100	0.017 0.346	-0.061 0.249	0.027 0.336	-0.123 0.086
Volume				1	<b>0.416</b> <b>0.002</b>	<b>0.505</b> <b>0.000</b>	0.202 0.120	<b>0.570</b> <b>0.000</b>	<b>0.511</b> <b>0.000</b>
Trade Size					1	0.167 0.085	0.205 0.070	0.165 0.098	0.166 0.092
Daily sd						1	<b>0.354</b> <b>0.042</b>	<b>0.928</b> <b>0.000</b>	<b>0.619</b> <b>0.000</b>
Tick sd							1	<b>0.366</b> <b>0.038</b>	<b>0.645</b> <b>0.000</b>
Daily RV								1	<b>0.686</b> <b>0.000</b>
Tick RV									1

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

**Figure 6.4:** An Illustration of Price Impact and VPIN from Samsung Electronics

We show an overlay of two time-series charts, one of price impact and one of VPIN. It can be seen on October 2007, price impact falls and VPIN rises. Also, in September 2008 price impact falls and VPIN rises.



### 6.6 Empirical Results

Here, we conduct a more robust empirical analysis on price impact and VPIN, as Pearson's correlations alone is insufficient in determining the relationship and possible lead lag structures that may exist. From simulations conducted on our theoretical model, we are aware that volume and volatility also play a part. Therefore, we propose to model the empirical relationship between price impact ( $\pi_t$ ), VPIN ( $VPIN_t$ ), volume ( $v_t$ ) and volatility<sup>1</sup> ( $\sigma_t$ ) using a VAR model. Our VAR model allows for both contemporaneous and lagged relationships.

$$\begin{aligned}
 \pi_t &= \sum_{i=1}^p \beta_{1,i}^{\pi} \pi_{t-i} + \sum_{i=0}^p \beta_{2,i}^{\pi} VPIN_{t-i} + \sum_{i=0}^p \beta_{3,i}^{\pi} v_{t-i} + \sum_{i=0}^p \beta_{4,i}^{\pi} \sigma_{t-i} + \varepsilon_{1,t} \\
 VPIN_t &= \sum_{i=0}^p \beta_{1,i}^{VPIN} \pi_{t-i} + \sum_{i=1}^p \beta_{2,i}^{VPIN} VPIN_{t-i} + \sum_{i=0}^p \beta_{3,i}^{VPIN} v_{t-i} + \sum_{i=0}^p \beta_{4,i}^{VPIN} \sigma_{t-i} + \varepsilon_{2,t} \\
 v_t &= \sum_{i=0}^p \beta_{1,i}^v \pi_{t-i} + \sum_{i=0}^p \beta_{2,i}^v VPIN_{t-i} + \sum_{i=1}^p \beta_{3,i}^v v_{t-i} + \sum_{i=0}^p \beta_{4,i}^v \sigma_{t-i} + \varepsilon_{3,t} \\
 \sigma_t &= \sum_{i=0}^p \beta_{1,i}^{\sigma} \pi_{t-i} + \sum_{i=0}^p \beta_{2,i}^{\sigma} VPIN_{t-i} + \sum_{i=0}^p \beta_{3,i}^{\sigma} v_{t-i} + \sum_{i=1}^p \beta_{4,i}^{\sigma} \sigma_{t-i} + \varepsilon_{4,t}
 \end{aligned} \tag{6.5}$$

The frequency of the VAR model is weekly, as this is the frequency of our time-varying price impact measure. We set  $p = 4$ , to reflect lags of up to 1 month. As our sample consists of weekly data from January 2007 to December 2012, we are aware that  $p$  cannot be too high to maintain a reasonable degree of freedom. Since there are 76 parameters in our VAR model, variable selection is necessary to achieve a sparse outcome. Using the same approach as in section 3.2

<sup>1</sup>Daily realized volatility

and 3.3, we apply regularization using Zou (2006) adaptive lasso. We feel lasso is suitable for this problem where parameters are dangerously high, as stated in Tibshirani (2013) it is even often used successfully in situations where parameters exceed observations. The variables are standardized (normalized) prior to estimation.

The results are consistent with our initial findings using Pearson's correlations and our theoretical model. For the majority of the stocks in our dataset we find a statistically significant contemporaneous negative correlation between price impact and VPIN. We also find no signs of significant lead-lag relationships between price impact and VPIN, and therefore it would be incorrect to suggest the possibility of Granger causality from VPIN to price impact or vice versa.

Consistent with our results in chapter 5, volatility is positively correlated with price impact. Whilst we note a positive coefficient between volume and price impact on some stocks, it was only present for a minority of stocks. However, VPIN seems to be unaffected by either volatility or volume. Both price impact and VPIN displayed positive autoregressive features. All four lags were significant for the majority of stocks with regards to VPIN. Two lags were significant for price impact.

Our results confirm the negative relationship between price impact and VPIN. We conclude that periods with low price impact provides opportunity for order imbalance to materialize. In scenarios where price impact is high, it is a lot harder for order imbalance to materialize as quote adjustments are high. We also confirm that if volatility is high, price impact is also high and VPIN is low.

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

**Table 6.3:** Adaptive Lasso VAR Summary

We tabulate the average significant coefficients from adaptive lasso VAR estimation on 24 individual Korean stocks. % sign. refers to the percentage of the 24 stocks that had showed significance for the specific coefficient using adaptive lasso regularization. The average adjusted  $R^2$  for each individual equation component of the VAR is also presented.

	Price Impact Equation			VPIN Equation			Volume Equation			Volatility Equation		
	lag	mean	% sig.	lag	mean	% sig.	lag	mean	% sig.	lag	mean	% sig.
Price Impact	0			<b>0</b>	<b>-0.160</b>	<b>79%</b>	0	0.001	17%	<b>0</b>	<b>0.223</b>	<b>92%</b>
	1	<b>0.412</b>	<b>100%</b>	1	-0.006	25%	1	-0.017	17%	1	0.022	25%
	2	<b>0.100</b>	<b>67%</b>	2	-0.002	13%	2	0.003	4%	2	-0.002	4%
	3	0.027	25%	3	-0.002	4%	3	-0.001	4%	3	-0.003	4%
	4	0.030	29%	4	-0.004	8%	4	0.003	4%	4	-0.002	4%
VPIN	<b>0</b>	<b>-0.072</b>	<b>54%</b>	0			0	0.025	33%	0	-0.002	4%
	1	-0.002	4%	<b>1</b>	<b>0.158</b>	<b>88%</b>	1	-0.002	4%	1	-0.002	4%
	2	0.000	0%	<b>2</b>	<b>0.060</b>	<b>50%</b>	2	-0.003	4%	2	0.000	0%
	3	0.000	8%	<b>3</b>	<b>0.073</b>	<b>71%</b>	3	0.000	0%	3	0.000	4%
	4	0.000	0%	<b>4</b>	<b>0.070</b>	<b>67%</b>	4	-0.013	13%	4	0.004	4%
Volume	0	0.019	17%	0	0.048	42%	0			<b>0</b>	<b>0.486</b>	<b>96%</b>
	1	0.007	8%	1	-0.015	21%	<b>1</b>	<b>0.372</b>	<b>100%</b>	<b>1</b>	<b>-0.117</b>	<b>63%</b>
	2	-0.001	8%	2	-0.023	33%	<b>2</b>	<b>0.072</b>	<b>58%</b>	2	-0.020	17%
	3	0.006	8%	3	-0.007	8%	3	0.051	42%	3	-0.026	25%
	4	0.002	13%	4	-0.006	13%	4	0.045	46%	4	-0.014	21%
Volatility	<b>0</b>	<b>0.180</b>	<b>92%</b>	0	-0.003	17%	<b>0</b>	<b>0.430</b>	<b>96%</b>	0		
	1	0.050	42%	1	-0.004	13%	1	-0.053	38%	<b>1</b>	<b>0.174</b>	<b>79%</b>
	2	0.014	25%	2	0.004	8%	2	-0.012	13%	2	0.034	29%
	3	0.002	4%	3	-0.004	8%	3	-0.016	29%	3	0.044	38%
	4	0.000	0%	4	0.000	0%	4	-0.010	8%	4	0.009	13%
Adj $R^2$	58.1%			24.0%			57.3%			53.9%		

## 6.7 Theoretical Explanation

Traditional intuition, would point to the fact that if information asymmetry existed (due to new information), then insiders would take advantage and trade for a profit. This would yield a greater order imbalance and also higher price impact as market makers and liquidity traders react to the incoming order flow. Certainly the works of Chung, Li and McInish (2005) show that with higher trading activity (a proxy of new information), larger price impact and stronger serial trade correlations appeared. Cross-sectionally, they found that higher price impact was positively related to PIN.

Therefore, it is almost counter-intuitive to suggest that a negative relationship exists between price impact and VPIN in time-series. Unfortunately, our empirical results in section 6.6 suggests this is the case. Therefore, there is a clear distinction in time-series results versus cross-sectional results.

Here we construct an information model to show that it is theoretically possible to have a negative relationship between price impact and VPIN with rational agents. Unlike information models based on Kyle (1985) or Easley and O'Hara (1992), we do not simply divide traders (agents) into the informed and uninformed. Instead we have a heterogeneous population of traders whose fundamental valuations are Normally distributed. When new information enters the market, we assume it slowly spreads through the population until every agent becomes informed (much like how a virus spreads through the human population). Therefore in our model, every agent begins as being uninformed and ends up being informed; and adjusts his or her valuations accordingly as he / she is updated with the new information. Also, unlike existing information models, we assume new information is constantly entering into the market, and before one piece of information is fully learned by the population, another piece has already entered. We believe this is more reflective of real-life trading, where there is always a constant stream of information for traders from telephones, email, television and Bloomberg.

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

We run simulations using these rational heterogeneous agents to show that when the underlying information flow becomes more variable (i.e., higher variance), price impact increases and VPIN decreases. We also find that volume and volatility both increase. Our volume-volatility result from our simulation is consistent with empirical findings by Karpoff (1987), Gallant, Rossi and Tauchen (1992) and Zhao and Wang (2003). Furthermore, when we run simulations by varying individual trader confidence levels (confidence as defined by Hong, Scheinkman and Xiong, 2006), we find that price impact increases but VPIN remains the same; and also volume decreases but volatility remains the same.

Firstly, our model explains the negative correlation between price impact and VPIN as documented in section 6.6. Secondly, it also shows that if trader confidence fluctuates, price impact moves accordingly irrespective of order flow imbalance and VPIN. This explains why figure 6.3 illustrates a significant but weak negative correlation. Suppose information flow variance is increasing, and with the increased variability trader confidence is decreasing, then our simulations show an increase in price impact and return volatility, and a decrease in VPIN. However, volume would remain uncertain as higher information variance increases it, but lower trader confidence decreases it. This matches the empirical results in table 6.3, where price impact is negatively correlated with VPIN, positively correlated with volatility, and not correlated with volume. In section below, we explain in detail the model we use to derive these relationships.

### 6.7.1 Heterogeneous agents and continuous information flow

Consider two points in time  $t = 0$  and  $t = 1$  and two assets namely a risky stock with price  $P_t$  and a riskfree bond with a riskfree rate of zero (for simplicity without lack of generality). Suppose investors decide on their allocation at time  $t = 0$  with  $X^f$  invested in the riskfree bond and  $X$  invested in the risky stock.

This implies, the investor's wealth at  $t = 0$  is  $W_0 = X^f + P_0X$  and at  $t = 1$  is  $W_1 = X^f + P_1X = W_0 + X(P_1 - P_0)$ . Furthermore, it is trivial to see that expected value and variance for  $W_1$  at time  $t = 0$  are  $E_0(W_1) = W_0 + X(E_0(P_1) - P_0)$  and  $Var_0(W_1) = X^2Var_0(P_1)$

respectively.

Let us assume that investors have constant absolute risk aversion (CARA) governed by a negative exponential utility function  $U(W_t) = -e^{-\rho W_t}$  with the Arrow-Pratt risk aversion coefficient denoted by  $\rho$ . This setup is standard with most rational expectations equilibrium models (see Grossman, 1976; Grossman and Stiglitz, 1980; Hussman, 1992; Romer, 1993; Baker and Stein, 2004 and Hong, Scheinkman and Xiong, 2006). With the assumption that wealth  $W$  is Gaussian distributed and subsequently employing its Gaussian moment generating function, the investor problem at  $t = 0$  is equivalent to maximizing,

$$aE_0(W_1) - \frac{\rho^2}{2}Var_0(W_1) \tag{6.6}$$

Solving for the first order condition yields the investor demand function,

$$X = \frac{E_0(P_1) - P_0}{\rho Var_0(P_1)} \tag{6.7}$$

This linear demand function is irrespective of initial wealth, and its a characteristic of the CARA assumption.

Let us create a heterogeneous market with  $n$  buyers and sellers. Let there be no short-selling. The demand function for buyers  $i = 1, \dots, n$  would be,

$$X_i^D = \frac{(V_i - P)_+}{\rho(\sigma_i^D)^2} \tag{6.8}$$

where  $V_i$  is the expected value of the stock at time  $t = 1$  given the information of trader  $i$  at time  $t = 0$ . Similarly,  $\sigma_i^2$  is the variance of the valuation of the stock at time  $t = 1$  given the information of trader  $i$  at time  $t = 0$ . We can view  $\frac{1}{\sigma_i^2}$  to be the information accuracy or confidence of the trader in his information signal. Likewise, the supply function for sells  $i = 1, \dots, n$  would be,

$$X_i^S = \min\left\{\frac{(P - V_i)_+}{\rho\sigma_i^2}, \omega_i\right\} \tag{6.9}$$

where  $\omega_i$  denotes the endowment in stock of each trader  $i$ . Here we do not allow short-selling, and thus traders can only sell what they have.



## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

Therefore, the aggregate demand and supply functions are,

$$X^D = \sum_{i=1}^n X_i^D, \quad X^S = \sum_{i=1}^n X_i^S \quad (6.10)$$

The market clearing conditions is  $X^D = X^S$  from which we can determine the equilibrium traded price and quantity.

Firstly, let us first consider a simple scenario where there exists a single piece of news information in the market with a tangible value of  $\eta$ . Without a loss of generality, let the news be positive such that  $\eta > 0$ . Therefore, prior to the news, the underlying price of the risky asset is  $p_0$  and as a consequence of the new information, the underlying price should be  $p_0 + \eta$ . However, it is unreasonable to assume that information dissemination affects all market participants at the same time, and therefore some traders receive the information before others. Let  $\lambda_t$  be the percentage of informed traders at time  $t$ . And let us now consider a time period from  $t = 0$  to  $T$ .

Our model is such that we have  $N$  market participants whose individual supply and demand functions are governed by CARA, none of which are informed  $\lambda_0 = 0$  at time  $t = 0$ . These uninformed traders have valuations drawn from the distribution,

$$V_{uninf} \sim \ln N(p_0, \sigma_{V_{uninf}}^2) \quad (6.11)$$

where  $\sigma_{V_{uninf}}$  is the dispersion parameter for the heterogeneity of uninformed trader beliefs.

At time  $t = T$ ,  $\lambda_T = 1$  and hence all the traders will be informed. Informed traders have valuations drawn from the distribution,

$$V_{inf} \sim \ln N(p_0 + \eta, \sigma_{V_{inf}}^2) \quad (6.12)$$

where  $\sigma_{V_{inf}}$  is the dispersion parameter for the heterogeneity of informed trader beliefs who have learned information signal  $\eta$ .

We assume that  $\lambda_t$  increases at a linear rate of  $1/T$ .

The initial endowment for our  $N$  traders  $\omega_{i,t=0}$  is distributed Exponentially,

$$\Omega \sim \text{Exp}(1/\omega) \quad (6.13)$$

where  $\omega$  is the mean endowment across the population.

If  $\rho$  and  $\sigma_i^2$  is heterogeneous between participants, then its effect is indistinguishable to valuation confidence  $\sigma_i^2$ , as they both impact the slopes of the individual demand and supply functions. Therefore risk aversion coefficient in our model is a market wide constant, and therefore we set  $\rho = 1$ . When  $\rho$  is homogeneous across all participants, it has no impact on equilibrium prices; however it does impact equilibrium volume. We assume individual confidence ( $\frac{1}{\sigma_i}$ ) is homogeneous amongst traders, i.e.  $\sigma_i^2 = \sigma_c^2, \forall i$ . This means that the individual demand and supply slopes of are the same. If the assumption is made that the individual confidence level is independent to the individual valuations (beliefs), then it is not unreasonable to simply take the average valuation confidence, if indeed there is some degree of heterogeneity in  $\sigma_i^2$ .

Consider the period  $0 \leq t \leq T$ , where  $0 \leq \lambda_t \leq 1$ .

Each individual informed supply curve for sellers  $i = 1, \dots, \lambda_t N$  at time  $t$  is,

$$X_{inf,i,t}^S = \min\left\{\frac{(v_{inf,i} - p_t)_+}{\rho\sigma_{inf,i}^2}, \omega_{i,t}\right\} \rightarrow \min\left\{\frac{(v_{inf,i} - p_t)_+}{\sigma_c^2}, \omega_{i,t}\right\} \quad (6.14)$$

the gradient of the function has been normalized; and  $v_{inf,i}$  is drawn from the Log-Normal r.v.  $V_{inf}$ .

Each individual uninformed supply curve for sellers  $i = \lambda_t N + 1, \dots, N$  at time  $t$  is,

$$X_{uninf,i,t}^S = \min\left\{\frac{(v_{uninf,i} - p_t)_+}{\rho\sigma_{uninf,i}^2}, \omega_{i,t}\right\} \rightarrow \min\left\{\frac{(v_{uninf,i} - p_t)_+}{\sigma_c^2}, \omega_{i,t}\right\} \quad (6.15)$$

the gradient of the function has been normalized; and  $v_{uninf,i}$  is drawn from the Log-Normal r.v.  $V_{uninf}$ .

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

Individual demand curves for informed sellers  $i = 1, \dots, \lambda_t N$  and uninformed sellers  $i = \lambda_t N + 1, \dots, N$  at time  $t$  are respectively,

$$X_{inf,i,t}^D = \frac{(p_t - v_{inf,i})_+}{\rho \sigma_{inf,i}^2} \rightarrow \frac{(p_t - v_{inf,i})_+}{\sigma_c^2} \quad (6.16)$$

$$X_{uninf,i,t}^D = \frac{(p_t - v_{uninf,i})_+}{\rho \sigma_{uninf,i}^2} \rightarrow \frac{(p_t - v_{uninf,i})_+}{\sigma_c^2} \quad (6.17)$$

Aggregate informed/uninformed supply/demand functions are  $X_{inf,t}^S = \sum_{i=1}^{\lambda_t N} X_{inf,i,t}^S$ ,  $X_{uninf,t}^S = \sum_{i=\lambda_t N+1}^N X_{uninf,i,t}^S$ ,  $X_{inf,t}^D = \sum_{i=1}^{\lambda_t N} X_{inf,i,t}^D$  and  $X_{uninf,t}^D = \sum_{i=\lambda_t N+1}^N X_{uninf,i,t}^D$ . Furthermore, total aggregate supply/demand are  $X_t^S = X_{inf,t}^S + X_{uninf,t}^S$  and  $X_t^D = X_{inf,t}^D + X_{uninf,t}^D$ .

Hence, traded price  $P_t$  and quantity  $Q_t$  can be determined by equating  $X_t^S = X_t^D$ . After determining  $P_t$ , we can work out the quantity traded by each trader by substituting  $P_t$  into each individual informed/uninformed supply/demand functions. For each individual trader  $i$ , we are able to work out whether they bought or sold, and subsequently can update their endowment  $\omega_{i,t}$  for the next period.

Price impact is estimated to be the inverse slope of the aggregate supply and demand functions at the equilibrium price  $P_t$ . This is because a unit buy trade shock will cause prices to go up by the gradient of the supply curve, and a unit sell trade shock will cause the prices to go down by the gradient of the demand curve. We use the delta approximation method to work out the slopes, and then take the average of the two. If we let price impact be  $\pi_t$ ,

$$\pi_t \approx \frac{1}{2} \left\{ \frac{X_t^S(P_t + \delta) - X_t^S(P_t - \delta)}{2\delta} + \frac{X_t^D(P_t - \delta) - X_t^D(P_t + \delta)}{2\delta} \right\} \quad (6.18)$$

To proxy VPIN, we consider the order flow of our model. Each trade consists of an active trader (who initiates the trade) and a passive trader who acts as the counter-party. We order all the traders from most informed to least informed (traders who first receive the information to traders who last receive the information). We then work out the net of all transactions (signed volume) of traders 1 to  $\frac{N}{2}$  and call it the net active trade volume ;and then net of all transactions of traders  $\frac{N}{2} + 1$  to  $N$  and call it the net passive trade volume. These two values should be the

identical in magnitude and opposite in sign. The absolute value of either one is the order flow imbalance. The rationale for this is as follows: (a) if the more informed traded amongst each other, then there should be 50-50 in terms of buy and sell initiations (b) if the less informed traded amongst each other, then there should also be 50-50 in terms of buy and sell initiation (c) only when the more informed trades with the less informed do we have order imbalance. Therefore, we divide the traders into two groups to determine order imbalance. VPIN is then estimated to be the imbalance volume divided by total traded volume.

**The Trading process:** Our heterogeneous trading process for  $t = 0, \dots, T$  periods where there is a single information flow is as follows:

1. Set  $\lambda_0 = 0$
2. Generate initial endowments  $\omega_i \quad \forall i = 1, \dots, N$ , from the r.v.  $\Omega$ .
3. Generate uninformed valuations  $v_{uninf,i} \quad \forall i = 1, \dots, N$ , from the r.v.  $V_{uninf}$ .
4. Generate informed valuations  $v_{inf,i} \quad \forall i = 1, \dots, N$ , from the r.v.  $V_{inf}$ .
5. Build functions  $X_{uninf,t=0}^S, X_{inf,t=0}^S, X_{uninf,t=0}^D, X_{inf,t=0}^D$
6. For  $t = 0$  to  $T$ :
  - . Equate  $X_{uninf,t}^S + X_{inf,t}^S = X_{uninf,t}^D + X_{inf,t}^D$  to work out  $(P_t, Q_t)$
  - . For  $i = 1$  to  $\lambda_t N$ 
    - . Quantity brought and sold by informed traders  $Q_{i,t} \leftarrow X_{inf,i,t}^D(P_t) - X_{inf,i,t}^S(P_t)$
  - . For  $i = \lambda_t N + 1$  to  $N$ 
    - . Quantity brought and sold by uninformed traders  $Q_{i,t} \leftarrow X_{uninf,i,t}^D(P_t) - X_{uninf,i,t}^S(P_t)$
  - . Update the endowments for the next period  $\omega_{i,t+1} \leftarrow \omega_{i,t} + Q_{i,t}$
  - . Update percentage informed  $\lambda_{t+1} \leftarrow \lambda_t + r$  by a learning rate  $r = 1/T$ .

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

---

. Update  $X_{uninf,t+1}^S, X_{inf,t+1}^S, X_{uninf,t+1}^D, X_{inf,t+1}^D$

It is unrealistic to assume that there is only a single piece of new information  $\eta$  being learned by market participants at a time. Here we consider a scenario where several information processes being learned by traders sequentially. Let  $1/\kappa$  be the linear speed at which the population learns of a single piece of information. Therefore it takes  $\kappa$  periods before a single piece of information is fully learnt, i.e.  $\lambda_\kappa = 1$ . Now we assume at every period, a new piece of information is introduced. Therefore at every given point in time, we have  $\kappa$  information processes that is being learned simultaneously by the trading population.

Denote  $\eta_t$  to be the new information that is disseminated into the population at time  $t$  with a learning rate of  $1/\kappa$ . At time  $t$ , these informations  $\eta_t, \eta_{t-1}, \dots, \eta_{t-\kappa+1}$  are still being learned. These information realizations are generated from the following distribution,

$$\eta \sim N(\mu_\eta, \sigma_\eta^2) \tag{6.19}$$

A restriction is the sequential nature at which the information is learned, i.e. a trader is required to learn  $\eta_{t-1}$  before he or she learns  $\eta_t$ .

In this simulation, we can adjust the underlying information process by adjusting  $\mu_\eta, \sigma_\eta^2$ , but leave the speed  $\kappa$  constant. Whilst we believe there is reason to believe that some time periods will consist of more price significant information or more variable information, we think there is little reason in changing the speed of information dissemination across time (unless there is some major technological breakthrough in communication, such as the invention of the Internet). We also adjust individual trader confidence  $\sigma_c^2$ , which we know will impact the slope of the aggregate supply and demand curves. Using the algorithm as shown, we run a series of simulations in R. Our objective is to show how heterogeneity impacts price discovery, volume, liquidity and volatility.

## 6.7 Theoretical Explanation

**Table 6.4:** Change of Information Flow Mean -  $\mu_\eta$

In this series of simulations we vary  $\mu_\eta$ , the mean valuation impact of the incoming information flow. We set the variance of the information flow  $\sigma_\eta^2 = 1$ . Individual trader valuation variance (the inverse of trader confidence) is  $\sigma_{inf}^2 = \sigma_{uninf}^2 = 1$ . Market wide risk aversion is  $\rho = 1$ . The speed of incoming information flow is  $\kappa = 3$ . We let there be  $n = 1000$  participants in the market, where the mean endowment  $\omega = 100$ . For each  $\mu_\eta$  setting, we take the average of 10 run, with each run consisting of 50 periods. Therefore, 5000 equilibriums have been generated for this table. The starting equilibrium price is set as \$10.

Information Mean	1st moment					2nd moment			
	Return	Volume	Price Impact	Imbal	VPIN	Return	Volume	Price Impact	Imbal
0.5	0.02	497	2.01E-03	158	32%	0.03	100	5.18E-06	224
1	0.04	574	2.01E-03	328	57%	0.03	140	1.02E-05	228
1.5	0.04	677	2.02E-03	482	71%	0.03	183	1.76E-05	229
2	0.05	814	2.02E-03	631	78%	0.04	227	2.56E-05	262
2.5	0.05	964	2.03E-03	733	76%	0.04	246	2.95E-05	279
3	0.06	1,109	2.03E-03	809	73%	0.04	269	2.95E-05	322
3.5	0.06	1,281	2.03E-03	897	70%	0.05	270	2.92E-05	362
4	0.06	1,419	2.03E-03	967	68%	0.05	291	3.26E-05	393
4.5	0.06	1,604	2.03E-03	1,059	66%	0.06	299	3.31E-05	435
5	0.07	1,761	2.03E-03	1,137	65%	0.06	318	3.63E-05	482

### 6.7.2 Simulated Results

Here we document our simulated results. In table 6.4 we document the change to equilibrium transactions, when we change the magnitude of the information flow ( $\mu_\eta$ ). When the new information flow is constantly large (in our simulation we used positive shocks), it is obvious that the return and the return standard deviation will increase. This is clearly shown by the first two movements of returns in table 6.4. We note that volume also increase. There the positive volume-volatility relationship (Karpoff, 1987) is evident here. Price impact increases as the information flow mean increase. This makes intuitive sense even in Kyle's (1985) model, where one would argue that the market markers are adjusting for greater levels of information. An interesting finding is that as both volume and imbalance volume (where more informed traders trade against less informed traders) are increasing, the ratio which we proxy as VPIN initially increases and then starts to decrease. This means that in periods where there is little information, there exists a positive relationship between VPIN and price impact. However, once we are in the region where the information is more significant we see a negative relationship.

In table 6.5 we change the information flow variance ( $\sigma_\eta^2$ ). From table 6.5, we find changing the variability of information flow has no impact on mean returns but increases return volatility. This is intuitive even without the simulations. From table 6.5, we can clearly see that

## 6. THE PRICE IMPACT OF A TRADE, VPIN AND THE ROLE OF INFORMED TRADER HETEROGENEITY

**Table 6.5:** Change of Information Flow Variance -  $\sigma_\eta^2$

In this series of simulations we vary  $\sigma_\eta^2$ , the variance of the incoming information flow. We set the information flow mean  $\mu_\eta = 5$ . Individual trader valuation variance (the inverse of trader confidence) is  $\sigma_{inf}^2 = \sigma_{uninf}^2 = 1$ . Market wide risk aversion is  $\rho = 1$ . The speed of incoming information flow is  $\kappa = 3$ . We let there be  $n = 1000$  participants in the market, where the mean endowment  $\omega = 100$ . For each  $\mu_\eta$  setting, we take the average of 10 run, with each run consisting of 50 periods. Therefore, 5000 equilibriums have been generated for this table. The starting equilibrium price is set as \$10.

Information Variance	1st moment					2nd moment			
	Return	Volume	Price Impact	Imbal	VPIN	Return	Volume	Price Impact	Imbal
1	0.07	1,796	2.04E-03	1,153	64%	0.06	340	3.95E-05	473
2	0.07	1,756	2.08E-03	1,117	64%	0.06	511	7.04E-05	532
3	0.07	1,877	2.11E-03	1,145	61%	0.07	761	8.45E-05	694
4	0.07	1,988	2.12E-03	1,169	59%	0.08	871	9.20E-05	788
5	0.07	2,075	2.13E-03	1,141	55%	0.10	1,064	9.38E-05	1,008
6	0.06	2,139	2.13E-03	1,113	52%	0.09	1,176	9.62E-05	1,217
7	0.06	2,273	2.14E-03	1,090	48%	0.20	1,285	9.72E-05	1,337
8	0.06	2,519	2.15E-03	1,078	43%	0.14	1,378	9.80E-05	1,570
9	0.07	2,891	2.17E-03	1,180	41%	0.12	1,594	9.27E-05	1,861
10	0.06	3,011	2.17E-03	1,134	38%	0.12	1,742	9.48E-05	1,884

price impact is increasing as information flow variance is increasing, whilst VPIN seems to be decreasing. Greater uncertainty in upcoming information reduces liquidity and increases price impact. This is not a new concept and uncertainty has long been associated with price impact; for example Ozsoylev and Werner (2011) utilize Ellsberg's paradox to explain how liquidity drops when ambiguous and uncertain information exists. As explained in Easley, Lopez de Prado and O'Hara (2011), we note a distinction between liquidity and volume. In this scenario, liquidity is declining, but volume is increasing due to greater information variance. We find that order imbalance remains indifferent to information flow volatility, but the overall volume is increasing. This causes the VPIN metric to decrease. In periods of greater uncertainty, it is not the number of trades between more informed and less informed traders that are increasing, it is the number of traders amongst the informed/uninformed groups that are increasing.

In table 6.6 we change individual trader confidence ( $\sigma_c^2$ ). By reducing the confidence levels of individual CARA governed traders, price impact increases. This is intuitive as traders become more risk averse. When risk aversion increases, we can see that the volume traded decreases significantly. In this scenario there is no changes to the underlying information flow, and so we find that returns, volatility and VPIN is constant. This experiment tells us that price impact is not only reflective of information content, as in pointed out by Hasbrouck (1991), but is also

## 6.7 Theoretical Explanation

**Table 6.6:** Change of Trader Confidence -  $\sigma_c^2$

In this series of simulations we vary  $\sigma_{inf}^2$ ,  $\sigma_{uninf}^2$ , the inverse of individual trader confidence. We set the information flow mean  $\mu_\eta = 2$  and variance  $\sigma_\eta^2 = 1$ . Market wide risk aversion is  $\rho = 1$ . The speed of incoming information flow is  $\kappa = 3$ . We let there be  $n = 1000$  participants in the market, where the mean endowment  $\omega = 100$ . For each  $\mu_\eta$  setting, we take the average of 10 run, with each run consisting of 50 periods. Therefore, 5000 equilibriums have been generated for this table. The starting equilibrium price is set as \$10.

(Lack of) Confidence	1st moment					2nd moment			
	Return	Volume	Price Impact	Imbal	VPIN	Return	Volume	Price Impact	Imbal
0.5	0.05	1,650	1.01E-03	1,308	79%	0.04	420	1.15E-05	484
1	0.05	814	2.02E-03	641	79%	0.04	226	2.55E-05	257
1.5	0.05	543	3.03E-03	428	79%	0.04	140	3.00E-05	164
2	0.05	402	4.04E-03	315	78%	0.03	109	4.42E-05	128
2.5	0.05	321	5.05E-03	252	78%	0.04	87	5.43E-05	100
3	0.05	272	6.07E-03	214	79%	0.04	74	6.26E-05	86
3.5	0.05	232	7.08E-03	183	79%	0.04	60	7.48E-05	71
4	0.05	206	8.09E-03	163	79%	0.03	56	9.09E-05	64
4.5	0.05	182	9.10E-03	143	79%	0.04	50	1.03E-04	59
5	0.05	162	1.01E-02	128	79%	0.03	41	1.08E-04	47

reflective of individual trader risk aversion.

Our model illustrates both heterogeneity amongst traders as well as information flow variability. We find the following results,

1. When information flow is changing in magnitude, price impact is positively correlated with volume and volatility. Initially with little information, there is a positive interaction between price impact and VPIN, but with greater information magnitude price impact is negatively related to VPIN
2. When information flow is changing in variance, price impact is positively correlated with volume and volatility and negatively correlated with VPIN
3. When trader confidence is changing, price impact is negatively correlated with volume. Volatility and VPIN is unchanged

Unless we are able to pin point exactly the latent information flow and trader confidence levels, we are never going to be able to full distinguish between these cases. What remains strictly true is that a general positive relationship between price impact and volatility and a general negative relationship between price impact and VPIN holds.



### 6.8 Conclusion

In conclusion, we have explained the construction of price impact and VPIN. We conducted weekly time-series analysis on both price impact and PIN and show empirically a negative correlation between the two measures exist. We show empirically and theoretically that information flow affects the two measures differently. Whilst risk aversion impacts price impact, it does not seem to impact VPIN. In Easley, Lopez de Prado and O'Hara (2011), it is found that VPIN is a good predictor for flash crashes. Their work was at the intra-day level. Our work shows the average VPIN at the weekly level is in fact negatively related to volatility and unrelated to volume. We argue that in periods where information flow volatility is high, price impact rises accordingly to accommodate for higher uncertainty, however VPIN declines due to the heterogeneous beliefs on the information. The greater variability of information means greater heterogeneity of valuations between informed traders (and hence more trading between them), which increases trading volume but not order imbalance.

# References

- [1] Abad, D. and J. Yague (2012) From PIN to VPIN: An introduction to order flow toxicity, *Spanish Review of Financial Economics* 10 74 - 83
- [2] Ane, T. and H. Geman (2000) Order flow, transaction clock, and normality of asset returns, *Journal of Finance* 55, 2259 - 2284
- [3] Baker, M. and J. Stein (2004) Market liquidity as a sentiment indicator, *Journal of Financial Markets* 7, 271 - 299
- [4] Brownlees, C.T. and G.M. Gallo (2006) Financial econometric analysis at ultra-high frequency: data handling concerns, *Computational Statistics & Data Analysis* 51, 2232 - 2245
- [5] Chakrabarty, B., Pascual, R. and A. Shkilko (2013) Trade Classification Algorithms: A Horse Race between the Bulk-Based and the Tick-Based Rules, working paper, Saint Louis University, USA
- [6] Chen, Y. and H. Zhao (2012) Informed trading, information uncertainty, and price momentum, *Journal of Banking & Finance* 36, 2095 - 2109
- [7] Chung, K.H., Li M. and T.H. McInish (2005) Information-based trading, price impact of trades, and trade autocorrelation, *Journal of Banking & Finance* 29, 1645 - 1669
- [8] Clark, P. (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices, *Econometrica* 41, 1351-55
- [9] Diamond, D.W. and R.E. Verrecchia (1987) Constraints on Short-Selling and Asset Price Adjustment to Private Information, *Journal of Financial Economics* 18, 277-311

## REFERENCES

---

- [10] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [11] Easley, D., Engle, R.F., OHara, M. and L. Wu (2008) Time-varying arrival rates of informed and uninformed trades, *Journal of Financial Econometrics* , 171 - 207
- [12] Easley, D., Hvidkjaer, S. and M. OHara, (2002) Is Information Risk a Determinant of Asset Returns?, *Journal of Finance* 10, 2185-2221
- [13] Easley, D., Kiefer, N.M. and M. O'Hara (1997) One day in the life of a very common stock, *The Review of Financial Studies* 10 - 3, 805 - 835
- [14] Easley, D., Kiefer, N., OHara, M., and J. Paperman, (1996) Liquidity, Information and Infrequently Traded Stocks, *Journal of Finance* 51, 1405-1436
- [15] Easley, D. and M.O'Hara (1992) Time and the process of security price adjustment, *Journal of Finance* 47, 577 -604
- [16] Easley, D., Lopez de Prado, M.M. and M. O'Hara (2011) The microstructure of the flash crash: flow toxicity, liquidity crashes, and the probability of informed trading, *Journal of Portfolio Management* 37, 118 - 128
- [17] Easley, D., Lopez de Prado, M.M. and M. O'Hara (2012) Flow toxicity and liquidity in a high-frequency world, *The Review for Financial Studies* 25, 1458 - 1493
- [18] Ellis, K., Michaely, R. and M. O'Hara (2000) The accuracy of trade classification rules: evidence from Nasdaq, *Journal of Financial and Quantitative Analysis* 35 - 4, 529 - 551
- [19] Escribano, A., Pascual., R. and M. Tapia (2004) Adverse selection costs, trading activity and price discovery in the NYSE: An empirical analysis *Journal of Banking & Finance* 28, 107 - 128
- [20] Foster, F.D. and S. Viswanathan (1993) Variations in trading volume, return volatility and trading costs: evidence on recent price formation models, *Journal of Finance* 48, 187 - 211

## REFERENCES

---

- [21] Gallant, A.R., Rossi, P.E. and G. Tauchen (1992) Stock Prices and Volume, *Review of Financial Studies* 5 - 2 199 - 242
- [22] Grossman, S (1980) On the efficiency of competitive stock markets where traders have diverse information, *Journal of Finance* 31 573 - 585
- [23] Grossman, S. and J.E. Stiglitz (1980) On the impossibility of informationally efficient markets, *The American Economic Review* 70 393 - 408
- [24] Hafner, C.M. (2005) Durations, volume and the prediction of financial returns in transaction time, *Quantitative Finance* 5 No.2, 145 - 152
- [25] Hasbrouck, J. (1991) Measuring the information content of stock trades, *Journal of Finance* 46, 179 - 208
- [26] Hasbrouck, J. (1991a) The summary informativeness of stock trades: an econometric analysis, *Review of Financial Studies* 4 No.3, 571 - 595
- [27] Hastie, T. and B. Efron (2011) Package 'lars', *CRAN repository*
- [28] Hsu, N.J., Hung, H.L. and Y.M. Chang (2008) Subset selection for vector autoregressive process using Lasso, *Computational Statistics and Data Analysis* 52, 3645 - 3657
- [29] Huang, R.D. and H.R. Stoll (1997) The components of the bid-ask spread: a general approach, *Review of Financial Studies* 10, 995 - 1034
- [30] Karpoff, J.M. (1987) The relation between price changes and trading volume: a survey, *Journal of Financial and Quantitative Analysis* 22 - 1, 109 - 126
- [31] Hong, H., Scheinkman, J. and W. Xiong (2006) Asset float and speculative bubbles, *Journal of Finance* 61, 1073 -1117
- [32] Jeria, D. and G. Sofianos (2008) Passive orders and natural adverse selection, *Street Smart* 33, Goldman Sachs
- [33] Kyle, A. P. (1985) Continuous Auctions and Insider Trading, *Econometrica* 53, 1315-1336.

## REFERENCES

---

- [34] Lee, C. and M. Ready (1991) Inferring trade direction from intraday data, *Journal of Finance* 46, 733 - 746
- [35] Lin, W.W. and W.C. Ke (2011) A computing bias in estimating the probability of informed trading, *Journal of Financial Markets* 14, 625 - 640
- [36] Manganello, S. (2005) Duration, volume and volatility impact of trades, *Journal of Financial Markets* 8, 377 - 399
- [37] Pesaran, H.H. and Y. Shin (1998) Generalized impulse response analysis in linear multivariate models, *Economics Letters* 58, 17 - 29
- [38] Ren, Y.W. and X.S. Zhang (2010) Subset selection for vector autoregressive processes via adaptive Lasso, *Statistics and Probability Letters* 80, 1705 - 1712
- [39] Romer, D. (1993) Rational asset-price movements without news, *The American Economic Review* 83, 1112 - 1130
- [40] Savin, I. (2010) A comparative study of the lasso-type and heuristic model selection models, *COMISEF working paper 042*, Max Planck Institute for Economics
- [41] Tay, A., Ting C., Tse Y.K. and M. Warachka (2009) Using high-frequency transactional data to estimate the probability of informed trading, *Journal of Financial Econometrics* 7 - 3, 288 - 311
- [42] Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society B* 58-1, 267-288
- [43] Tibshirani, R. (2013) The lasso problem and uniqueness, *Electronic Journal of Statistics* 7, 1456 - 1490
- [44] Werner, J. and H. Ozsoylev (2011) Liquidity and asset prices in rational expectations equilibrium with ambiguous information, *Economic Theory* 43, 469 - 491
- [45] Zhou B. (1996) Forecasting foreign exchange rates subject to devolatilization , chapter 3, 51-67, in C. Dunis [ed.], *Forecasting Financial Markets*, John Wiley & Sons, Chichester.

## REFERENCES

---

- [46] Zou, H. (2006) The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association* 101 No.476, 1418 - 1429

## REFERENCES

---

# 7

## Conclusions

In conclusion, we have presented empirical methodologies and analyzed relationships in the fields of information asymmetry and price impact in market microstructure. In particular, we focus on developing a new methodology for estimating the information asymmetry measure PIN (Easley, Kiefer, O'Hara and Paperman, 1996; Easley, Hvidkjaer and O'Hara, 2002, 2010), and showed the usage of adaptive lasso regularization in VAR models for estimating the price impact of a trade (Hasbrouck, 1991; Dufour and Engle, 2000). Using our newly developed adaptive lasso VAR model, we document the relationships between price impact and volatility, and also price impact and VPIN (Easley, Lopez de Prado and O'Hara, 2012). We also provide theoretical models to explain our empirical findings.

In our first paper (chapter 3), we show how hierarchical agglomerative clusters can be employed to estimate PIN and its components. We find that our new methodology to be comparable in terms of accuracy and 300x faster than the best existing method in the literature, i.e., Lin and Ke's (2011) method with Yan and Zhang's (2012) starting value algorithm. We also show that clusters allows for explicit classification of trading days into 'good news', 'bad news' and 'no news'. This cannot be achieved in traditional maximum likelihood methods and allows us to conduct ex-post analysis and validation of the classification. We test our cluster classifications to simple trading measures, such as returns, volatility and volume, and show the classifications to be sensible. Following from explicit classification, we can also map the flow of news events



## 7. CONCLUSIONS

---

and show it has significant autoregressive properties.

In our second paper (chapter 4), we extend upon Hasbrouck (199) and Dufour and Engle (2000) and show that order book illiquidity parameters can also impact quote revisions. We show our findings to be robust by utilizing a sparse VAR model governed by adaptive lasso regularization. We also find that contrary to Dufour and Engle (2000), periods with higher levels of trading tend to have lower price impact and vice versa.

By estimating price impact across time using the our adaptive lasso VAR model, we conduct time-series analysis between price impact and volatility. In our third paper (chapter 5), we show price impact and volatility to be positively correlated. We explain this using a rational expectations equilibrium model with a linear price rule.

Lastly (chapter 6), we examine the time-series relationship between price impact and VPIN. Price impact relates to the level liquidity traders revise their quotes due to incoming order flow, similar to Kyle's (1985)  $\lambda$ . VPIN is related to the order flow imbalance as percentage of volume. Both measures relate to illiquidity, risk aversion costs and information asymmetry, yet we find that they are indeed negatively correlated. Whilst this may seem counter-intuitive, we show that if informed traders have heterogeneous beliefs on the value of incoming information flow, then it is possible for price impact to be inversely related to VPIN.

Therefore, we have studied about information asymmetry and price impact in market microstructure. We have provided some our own additions to the literature in terms of methodology (i.e., CPIN and adaptive lasso VAR for price impact) and subsequently explored some new relationships ( i.e., price impact and volatility, and price impact and VPIN).

# References

- [1] Dufour, A. and R.F. Engle (2000) Time and the price impact of a trade, *Journal of Finance* 55, 2467 - 2498
- [2] Easley, D., Hvidkjaer, S. and M. OHara, (2002) Is Information Risk a Determinant of Asset Returns?, *Journal of Finance* 10, 2185-2221
- [3] Easley, D., Hvidkjaer, S. and M. OHara, (2010) Factoring information into returns, *Journal of Financial and Quantitative Analysis* 45 - 2, 293 - 309
- [4] Easley, D., Kiefer, N., OHara, M., and J. Paperman, (1996) Liquidity, Information and Infrequently Traded Stocks, *Journal of Finance* 51, 1405-1436
- [5] Hasbrouck, J. (1991) Measuring the information content of stock trades, *Journal of Finance* 46, 179 - 208
- [6] Kyle, A.S. (1985) Continuous auctions and insider trading, *Econometrica* 53 - 6, 1315 - 1336
- [7] Lin, W.W. and W.C. Ke (2011) A computing bias in estimating the probability of informed trading, *Journal of Financial Markets* 14, 625 - 640
- [8] Yan, Y. and S. Zhang (2012) An improved estimation method and empirical properties of the probability of informed trading, *Journal of Banking & Finance* 36, 454 - 467

## REFERENCES

---

## Appendix A

# Supplement for Chapter 4

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.1:** Adaptive Lasso Coefficients Samsung Electronics (005930 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value	***	lag	mean	stdev	T-stat	p-value	***
Quote Revisions	1	<b>-0.097</b>	0.070	-24.55	0.000	***	1	<b>9.857</b>	4.605	37.87	0.000	***
	2	<b>-0.065</b>	0.064	-18.09	0.000	***	2	<b>-9.513</b>	3.974	-42.35	0.000	***
	3	<b>-0.039</b>	0.048	-14.39	0.000	***	3	<b>-5.350</b>	2.423	-39.06	0.000	***
	4	<b>-0.032</b>	0.050	-11.17	0.000	***	4	<b>-4.448</b>	1.905	-41.31	0.000	***
	5	<b>-0.023</b>	0.037	-10.79	0.000	***	5	<b>-3.157</b>	1.524	-36.65	0.000	***
	6	<b>-0.016</b>	0.032	-8.79	0.000	***	6	<b>-2.289</b>	1.329	-30.48	0.000	***
	7	<b>-0.012</b>	0.031	-7.12	0.000	***	7	<b>-1.621</b>	1.109	-25.87	0.000	***
	8	<b>-0.007</b>	0.018	-6.80	0.000	***	8	<b>-1.063</b>	0.966	-19.47	0.000	***
	9	<b>-0.007</b>	0.017	-6.92	0.000	***	9	<b>-0.599</b>	0.710	-14.91	0.000	***
	10	<b>-0.006</b>	0.018	-5.85	0.000	***	10	<b>-0.259</b>	0.437	-10.49	0.000	***
Trades	0	<b>8.92E-04</b>	5.56E-04	28.39	0.000	***	1	<b>0.210</b>	0.078	47.91	0.000	***
	1	<b>2.68E-04</b>	2.13E-04	22.28	0.000	***	2	<b>0.066</b>	0.031	37.54	0.000	***
	2	<b>2.17E-04</b>	1.44E-04	26.68	0.000	***	3	<b>0.066</b>	0.013	89.30	0.000	***
	3	<b>1.40E-04</b>	1.37E-04	18.03	0.000	***	4	<b>0.052</b>	0.011	83.74	0.000	***
	4	<b>1.06E-04</b>	1.22E-04	15.39	0.000	***	5	<b>0.043</b>	0.010	76.86	0.000	***
	5	<b>6.87E-05</b>	9.86E-05	12.33	0.000	***	6	<b>0.038</b>	0.010	69.44	0.000	***
	6	<b>4.73E-05</b>	1.01E-04	8.32	0.000	***	7	<b>0.033</b>	0.010	60.53	0.000	***
	7	<b>3.21E-05</b>	6.39E-05	8.90	0.000	***	8	<b>0.031</b>	0.009	60.17	0.000	***
	8	<b>2.54E-05</b>	6.49E-05	6.92	0.000	***	9	<b>0.028</b>	0.010	50.73	0.000	***
	9	<b>2.18E-05</b>	6.69E-05	5.77	0.000	***	10	<b>0.030</b>	0.009	57.92	0.000	***
Durations	0	<b>2.49E-06</b>	2.21E-06	19.94	0.000	***	1	5.22E-06	7.95E-05	1.16	0.247	
	1	-9.08E-08	4.70E-06	-0.34	0.733		2	-4.69E-06	9.02E-05	-0.92	0.358	
	2	-2.70E-07	5.56E-06	-0.86	0.391		3	2.07E-06	7.52E-05	0.49	0.627	
	3	8.48E-08	5.24E-06	0.29	0.775		4	2.35E-06	8.69E-05	0.48	0.633	
	4	2.36E-07	6.61E-06	0.63	0.528		5	2.65E-06	5.33E-05	0.88	0.379	
	5	-1.82E-08	2.70E-06	-0.12	0.905		6	-6.30E-07	2.84E-05	-0.39	0.695	
	6	-6.51E-09	2.96E-06	-0.04	0.969		7	-5.41E-07	4.68E-05	-0.20	0.838	
	7	-2.45E-08	2.85E-06	-0.15	0.879		8	-2.06E-06	6.18E-05	-0.59	0.557	
	8	-1.60E-07	3.21E-06	-0.88	0.378		9	1.04E-06	3.81E-05	0.48	0.630	
	9	-1.06E-08	2.82E-06	-0.07	0.947		10	-3.79E-06	4.12E-05	-1.62	0.105	
Spreads	0	<b>3.15E-06</b>	1.47E-05	3.80	0.000	***	1	-5.05E-06	3.88E-04	-0.23	0.818	
	1	<b>1.94E-06</b>	1.11E-05	3.08	0.002	***	2	-7.95E-07	1.42E-04	-0.10	0.921	
	2	<b>1.36E-06</b>	9.08E-06	2.65	0.009	***	3	-3.87E-06	1.30E-04	-0.53	0.599	
	3	1.21E-06	1.05E-05	2.04	0.042	**	4	6.02E-06	7.72E-05	1.38	0.168	
	4	1.09E-06	7.47E-06	2.57	0.011	**	5	9.88E-07	7.42E-05	0.24	0.814	
	5	-5.33E-08	6.22E-06	-0.15	0.880		6	8.07E-06	6.22E-05	2.30	0.022	**
	6	-1.79E-07	4.98E-06	-0.64	0.525		7	4.26E-06	4.08E-05	1.84	0.066	*
	7	3.41E-07	5.34E-06	1.13	0.259		8	-1.96E-06	4.24E-05	-0.82	0.415	
	8	1.53E-07	5.37E-06	0.51	0.614		9	2.28E-06	4.93E-05	0.82	0.415	
	9	2.39E-07	4.43E-06	0.96	0.339		10	1.18E-06	6.77E-05	0.31	0.759	
Depth	0	<b>-7.47E-07</b>	5.87E-07	-22.53	0.000	***	1	<b>-1.10E-05</b>	1.40E-05	-13.86	0.000	***
	1	<b>-2.21E-07</b>	3.47E-07	-11.24	0.000	***	2	<b>-3.97E-06</b>	8.80E-06	-8.00	0.000	***
	2	<b>-1.07E-07</b>	1.99E-07	-9.54	0.000	***	3	<b>-1.97E-06</b>	5.67E-06	-6.15	0.000	***
	3	<b>-7.33E-08</b>	1.52E-07	-8.50	0.000	***	4	<b>-1.24E-06</b>	5.95E-06	-3.69	0.000	***
	4	<b>-4.24E-08</b>	1.12E-07	-6.73	0.000	***	5	<b>-8.69E-07</b>	3.24E-06	-4.75	0.000	***
	5	<b>-2.90E-08</b>	1.32E-07	-3.89	0.000	***	6	-3.44E-07	2.76E-06	-2.20	0.028	**
	6	<b>-3.72E-08</b>	1.88E-07	-3.51	0.001	***	7	7.94E-08	2.68E-06	0.52	0.601	
	7	-9.27E-09	1.33E-07	-1.23	0.219		8	3.72E-07	3.66E-06	1.80	0.073	*
	8	-7.06E-09	1.14E-07	-1.10	0.273		9	2.79E-07	3.00E-06	1.64	0.102	
	9	-3.44E-09	8.55E-08	-0.71	0.477		10	5.97E-07	3.40E-06	1.11	0.272	

**Table A.2:** Adaptive Lasso Coefficients Hyundai Motor (005380 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.090	0.060	-26.54	0.000	***	1	9.349	5.331	31.02	0.000	***
	2	-0.057	0.041	-24.54	0.000	***	2	-12.831	6.734	-33.71	0.000	***
	3	-0.038	0.027	-24.67	0.000	***	3	-6.523	2.825	-40.85	0.000	***
	4	-0.026	0.024	-18.73	0.000	***	4	-5.610	2.720	-36.48	0.000	***
	5	-0.017	0.016	-18.94	0.000	***	5	-3.890	1.971	-34.91	0.000	***
	6	-0.012	0.017	-12.39	0.000	***	6	-2.878	1.566	-32.51	0.000	***
	7	-0.008	0.015	-9.71	0.000	***	7	-1.916	1.267	-26.77	0.000	***
	8	-0.006	0.013	-8.06	0.000	***	8	-1.224	0.947	-22.86	0.000	***
	9	-0.004	0.010	-6.58	0.000	***	9	-0.647	0.776	-14.75	0.000	***
	10	-0.004	0.015	-4.42	0.000	***	10	-0.264	0.541	-8.64	0.000	***
Trades	0	7.89E-04	5.35E-04	26.06	0.000	***	1	0.234	0.081	50.88	0.000	***
	1	1.99E-04	1.69E-04	20.84	0.000	***	2	0.067	0.045	26.28	0.000	***
	2	1.72E-04	1.15E-04	26.47	0.000	***	3	0.073	0.015	85.65	0.000	***
	3	9.65E-05	9.41E-05	18.14	0.000	***	4	0.055	0.014	69.11	0.000	***
	4	6.16E-05	7.58E-05	14.37	0.000	***	5	0.047	0.012	67.68	0.000	***
	5	4.96E-05	7.43E-05	11.81	0.000	***	6	0.040	0.012	59.57	0.000	***
	6	3.22E-05	5.66E-05	10.06	0.000	***	7	0.036	0.012	54.68	0.000	***
	7	2.13E-05	4.51E-05	8.37	0.000	***	8	0.033	0.012	48.26	0.000	***
	8	1.37E-05	4.07E-05	5.96	0.000	***	9	0.030	0.013	41.20	0.000	***
	9	1.40E-05	4.17E-05	5.95	0.000	***	10	0.034	0.012	48.24	0.000	***
10	7.60E-06	2.45E-05	5.48	0.000	***							
Durations	0	2.99E-06	2.61E-06	20.29	0.000	***	1	2.15E-06	5.88E-05	0.65	0.519	
	1	3.52E-07	2.32E-06	2.68	0.008	***	2	4.62E-06	5.20E-05	1.57	0.117	
	2	2.39E-08	4.14E-06	0.10	0.919		3	9.17E-06	6.30E-05	2.57	0.011	**
	3	1.13E-07	2.49E-06	0.80	0.422		4	1.74E-06	7.81E-05	0.39	0.694	
	4	-1.69E-07	2.05E-06	-1.46	0.146		5	5.42E-06	4.54E-05	2.11	0.036	**
	5	-1.18E-07	3.53E-06	-0.59	0.554		6	-1.24E-06	5.04E-05	-0.44	0.663	
	6	1.46E-07	1.95E-06	1.32	0.187		7	-3.78E-06	7.49E-05	-0.89	0.372	
	7	4.18E-08	1.48E-06	0.50	0.617		8	-8.07E-07	3.65E-05	-0.39	0.696	
	8	-1.39E-07	1.65E-06	-1.50	0.135		9	1.64E-06	2.48E-05	1.17	0.242	
	9	8.16E-08	2.39E-06	0.60	0.547		10	-2.07E-06	3.62E-05	-1.01	0.314	
10	1.48E-07	2.09E-06	1.25	0.212								
Spreads	0	2.44E-05	6.64E-05	6.49	0.000	***	1	3.25E-04	1.94E-03	2.96	0.003	***
	1	1.44E-05	4.50E-05	5.66	0.000	***	2	1.37E-04	9.82E-04	2.47	0.014	**
	2	5.85E-06	4.36E-05	2.37	0.019	**	3	1.42E-04	1.11E-03	2.26	0.025	**
	3	3.10E-06	3.87E-05	1.41	0.159		4	-6.89E-06	6.32E-04	-0.19	0.848	
	4	1.53E-07	2.55E-05	0.11	0.916		5	1.14E-04	7.78E-04	2.59	0.010	***
	5	1.59E-06	2.25E-05	1.25	0.212		6	2.41E-05	3.45E-04	1.23	0.218	
	6	2.91E-06	2.87E-05	1.79	0.074	*	7	1.85E-05	4.65E-04	0.70	0.482	
	7	-5.18E-07	2.68E-05	-0.34	0.733		8	1.55E-05	3.80E-04	0.72	0.471	
	8	-3.19E-07	2.15E-05	-0.26	0.794		9	-7.20E-07	2.23E-04	-0.06	0.955	
	9	3.17E-06	2.68E-05	2.09	0.037	**	10	1.77E-06	3.61E-04	0.09	0.931	
10	9.88E-08	2.23E-05	0.08	0.938								
Depth	0	-2.98E-07	2.74E-07	-19.29	0.000	***	1	-7.26E-06	7.87E-06	-16.32	0.000	***
	1	-8.12E-08	1.06E-07	-13.59	0.000	***	2	-2.19E-06	5.30E-06	-7.33	0.000	***
	2	-4.72E-08	7.47E-08	-11.18	0.000	***	3	-1.15E-06	3.04E-06	-6.72	0.000	***
	3	-2.65E-08	5.06E-08	-9.28	0.000	***	4	-5.20E-07	3.42E-06	-2.69	0.008	***
	4	-1.32E-08	3.83E-08	-6.10	0.000	***	5	1.45E-08	3.34E-06	0.08	0.939	
	5	-6.57E-09	2.77E-08	-4.20	0.000	***	6	-1.29E-07	1.99E-06	-1.15	0.252	
	6	-3.33E-09	2.35E-08	-2.51	0.012	**	7	2.57E-07	2.74E-06	1.66	0.099	*
	7	-1.52E-09	1.94E-08	-1.39	0.166		8	2.75E-07	2.76E-06	1.76	0.080	**
	8	-1.35E-09	1.39E-08	-1.72	0.087	*	9	4.98E-07	2.80E-06	3.15	0.002	***
	9	1.13E-09	2.20E-08	0.91	0.364		10	7.84E-07	3.08E-06	1.10	0.272	
10	1.45E-09	1.58E-08	1.62	0.106								

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.3:** Average Adaptive Lasso Coefficients POSCO (005490 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value		lag	mean	stdev	T-stat	p-value	
Quote Revisions	1	-0.103	0.060	-30.51	0.000	***	1	10.125	3.618	49.50	0.000	***
	2	-0.060	0.052	-20.52	0.000	***	2	-8.222	3.334	-43.63	0.000	***
	3	-0.033	0.033	-17.66	0.000	***	3	-3.570	2.110	-29.94	0.000	***
	4	-0.023	0.032	-12.54	0.000	***	4	-2.833	1.727	-29.03	0.000	***
	5	-0.012	0.023	-9.72	0.000	***	5	-1.716	1.480	-20.51	0.000	***
	6	-0.010	0.020	-8.73	0.000	***	6	-1.126	1.135	-17.55	0.000	***
	7	-0.006	0.017	-5.68	0.000	***	7	-0.638	0.873	-12.93	0.000	***
	8	-0.006	0.019	-5.16	0.000	***	8	-0.404	0.686	-10.42	0.000	***
	9	-0.002	0.012	-3.36	0.001	***	9	-0.175	0.412	-7.54	0.000	***
	10	-0.002	0.012	-2.62	0.009	***	10	-0.098	0.299	-5.81	0.000	***
			mean	stdev								
Trades	0	1.33E-03	8.49E-04	27.65	0.000	***	1	0.221	0.077	50.96	0.000	***
	1	3.58E-04	2.89E-04	21.96	0.000	***	2	0.060	0.036	29.83	0.000	***
	2	2.51E-04	1.91E-04	23.23	0.000	***	3	0.061	0.013	83.87	0.000	***
	3	1.37E-04	1.48E-04	16.33	0.000	***	4	0.044	0.012	66.91	0.000	***
	4	7.36E-05	1.13E-04	11.50	0.000	***	5	0.037	0.011	58.88	0.000	***
	5	5.56E-05	1.07E-04	9.19	0.000	***	6	0.031	0.012	46.02	0.000	***
	6	3.61E-05	8.50E-05	7.51	0.000	***	7	0.027	0.011	42.13	0.000	***
	7	2.17E-05	5.45E-05	7.05	0.000	***	8	0.024	0.012	36.26	0.000	***
	8	1.29E-05	4.38E-05	5.22	0.000	***	9	0.022	0.011	34.42	0.000	***
	9	9.54E-06	4.39E-05	3.84	0.000	***	10	0.025	0.011	39.75	0.000	***
	10	6.06E-06	2.94E-05	3.65	0.000	***						
		mean	stdev									
Durations	0	3.54E-06	3.28E-06	19.11	0.000	***	1	9.34E-06	8.47E-05	1.95	0.052	*
	1	-3.34E-07	3.48E-06	-1.70	0.091	*	2	6.75E-06	5.65E-05	2.11	0.035	**
	2	-1.12E-07	3.14E-06	-0.63	0.526		3	3.58E-07	4.58E-05	0.14	0.890	
	3	-3.70E-08	3.92E-06	-0.17	0.867		4	1.61E-06	5.52E-05	0.51	0.607	
	4	-1.25E-07	3.06E-06	-0.72	0.470		5	9.50E-07	4.88E-05	0.34	0.731	
	5	-7.64E-08	1.74E-06	-0.78	0.438		6	-1.66E-06	3.00E-05	-0.98	0.328	
	6	-2.10E-08	1.51E-06	-0.25	0.806		7	-7.84E-07	2.21E-05	-0.63	0.530	
	7	-1.40E-07	2.20E-06	-1.13	0.259		8	2.83E-06	2.39E-05	2.10	0.037	**
	8	-1.56E-07	1.94E-06	-1.42	0.158		9	-8.63E-07	1.75E-05	-0.87	0.383	
	9	-3.53E-07	2.00E-06	-3.13	0.002	***	10	-7.36E-07	2.45E-05	-0.53	0.596	
	10	-7.45E-08	1.41E-06	-0.93	0.351							
		mean	stdev									
Spreads	0	3.87E-06	1.43E-05	4.78	0.000	***	1	1.86E-05	2.55E-04	1.29	0.197	
	1	2.56E-06	1.27E-05	3.57	0.000	***	2	-2.24E-06	1.33E-04	-0.30	0.766	
	2	2.29E-06	8.50E-06	4.78	0.000	***	3	-4.35E-06	1.17E-04	-0.66	0.510	
	3	3.80E-07	9.95E-06	0.68	0.499		4	1.28E-06	8.84E-05	0.26	0.798	
	4	8.20E-08	9.11E-06	0.16	0.874		5	-1.10E-06	5.22E-05	-0.37	0.710	
	5	8.29E-07	4.27E-06	3.44	0.001	***	6	-1.11E-07	3.55E-05	-0.06	0.956	
	6	3.49E-07	3.53E-06	1.75	0.081	*	7	3.95E-06	3.76E-05	1.86	0.064	*
	7	3.61E-07	4.54E-06	1.41	0.161		8	-1.39E-06	5.29E-05	-0.46	0.643	
	8	7.65E-09	4.20E-06	0.03	0.974		9	-1.07E-06	3.68E-05	-0.51	0.608	
	9	7.29E-09	3.03E-06	0.04	0.966		10	-1.56E-06	2.33E-05	-1.18	0.238	
	10	3.25E-07	2.83E-06	2.03	0.043	**						
		mean	stdev									
Depth	0	-1.56E-06	1.30E-06	-21.26	0.000	***	1	-1.17E-05	1.81E-05	-11.40	0.000	***
	1	-3.71E-07	4.91E-07	-13.35	0.000	***	2	-2.27E-06	9.51E-06	-4.23	0.000	***
	2	-1.70E-07	3.26E-07	-9.24	0.000	***	3	-7.41E-07	7.29E-06	-1.80	0.073	*
	3	-7.33E-08	3.23E-07	-4.01	0.000	***	4	6.46E-07	8.27E-06	1.38	0.168	
	4	-5.81E-08	3.76E-07	-2.73	0.007	***	5	1.03E-06	8.55E-06	2.12	0.035	**
	5	-1.72E-08	1.89E-07	-1.61	0.109		6	5.22E-07	4.37E-06	2.11	0.036	**
	6	5.19E-09	1.28E-07	0.72	0.474		7	8.03E-07	4.54E-06	3.13	0.002	***
	7	-8.14E-09	2.05E-07	-0.70	0.482		8	7.69E-07	4.58E-06	2.97	0.003	***
	8	-1.77E-08	2.46E-07	-1.27	0.205		9	8.98E-07	4.61E-06	3.44	0.001	***
	9	1.57E-08	2.79E-07	1.00	0.319		10	9.56E-07	4.39E-06	1.10	0.272	
	10	-3.86E-10	3.09E-07	-0.02	0.982							

**Table A.4:** Average Adaptive Lasso Coefficients Hyundai Mobis (012330 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value	***	lag	mean	stdev	T-stat	p-value	***
Quote Revisions	1	-0.123	0.060	-36.03	0.000	***	1	9.893	4.547	38.49	0.000	***
	2	-0.060	0.037	-28.46	0.000	***	2	-10.787	6.078	-31.40	0.000	***
	3	-0.041	0.028	-25.75	0.000	***	3	-4.089	2.051	-35.26	0.000	***
	4	-0.022	0.020	-19.17	0.000	***	4	-3.288	1.727	-33.69	0.000	***
	5	-0.014	0.028	-9.13	0.000	***	5	-1.813	1.295	-24.77	0.000	***
	6	-0.010	0.023	-7.54	0.000	***	6	-1.113	1.066	-18.47	0.000	***
	7	-0.005	0.017	-5.58	0.000	***	7	-0.644	0.818	-13.94	0.000	***
	8	-0.004	0.026	-2.79	0.006	***	8	-0.349	0.588	-10.51	0.000	***
	9	-0.001	0.013	-1.58	0.115	***	9	-0.168	0.441	-6.74	0.000	***
	10	-0.002	0.010	-2.77	0.006	***	10	-0.093	0.278	-5.93	0.000	***
Trades	0	1.49E-03	9.23E-04	28.47	0.000	***	1	0.256	0.096	47.34	0.000	***
	1	3.50E-04	2.95E-04	20.99	0.000	***	2	0.066	0.043	26.86	0.000	***
	2	2.82E-04	2.13E-04	23.38	0.000	***	3	0.068	0.015	81.16	0.000	***
	3	1.42E-04	1.57E-04	16.02	0.000	***	4	0.048	0.016	53.27	0.000	***
	4	8.17E-05	1.19E-04	12.13	0.000	***	5	0.039	0.014	49.70	0.000	***
	5	5.60E-05	9.51E-05	10.41	0.000	***	6	0.033	0.015	39.76	0.000	***
	6	3.42E-05	8.51E-05	7.11	0.000	***	7	0.026	0.014	32.78	0.000	***
	7	2.39E-05	7.12E-05	5.95	0.000	***	8	0.025	0.014	33.07	0.000	***
	8	1.59E-05	4.90E-05	5.74	0.000	***	9	0.024	0.013	32.32	0.000	***
	9	6.97E-06	3.69E-05	3.34	0.001	***	10	0.026	0.013	35.23	0.000	***
10	3.27E-06	2.44E-05	2.37	0.018	**							
Durations	0	5.03E-06	4.47E-06	19.91	0.000	***	1	-3.26E-08	8.48E-05	-0.01	0.995	
	1	-2.54E-07	4.10E-06	-1.10	0.273	2	4.78E-06	6.79E-05	1.25	0.214		
	2	1.18E-07	4.01E-06	0.52	0.603	3	6.96E-07	4.96E-05	0.25	0.804		
	3	-4.44E-08	3.99E-06	-0.20	0.844	4	4.05E-06	3.86E-05	1.85	0.065		
	4	1.65E-07	2.11E-06	1.38	0.167	5	1.95E-07	3.55E-05	0.10	0.923		
	5	-1.13E-07	3.50E-06	-0.57	0.568	6	2.43E-06	4.00E-05	1.07	0.284		
	6	6.97E-10	1.61E-06	0.01	0.994	7	7.95E-07	2.66E-05	0.53	0.597		
	7	4.40E-07	4.96E-06	1.57	0.117	8	-3.06E-06	5.11E-05	-1.06	0.290		
	8	-1.13E-07	2.06E-06	-0.98	0.330	9	1.68E-06	3.59E-05	0.83	0.408		
	9	1.51E-08	1.24E-06	0.22	0.829	10	4.48E-07	2.59E-05	0.31	0.760		
10	8.89E-09	1.52E-06	0.10	0.918								
Spreads	0	1.88E-05	4.10E-05	8.11	0.000	***	1	1.86E-04	1.08E-03	3.04	0.003	***
	1	8.41E-06	2.74E-05	5.42	0.000	***	2	9.13E-05	4.62E-04	3.49	0.001	***
	2	5.42E-06	2.13E-05	4.50	0.000	***	3	2.64E-05	2.50E-04	1.87	0.063	*
	3	2.93E-06	1.87E-05	2.77	0.006	***	4	2.75E-05	1.96E-04	2.48	0.014	**
	4	5.24E-07	2.32E-05	0.40	0.689	5	2.05E-05	2.91E-04	1.25	0.213		
	5	2.13E-06	1.97E-05	1.91	0.057	*	6	-8.27E-06	3.34E-04	-0.44	0.662	
	6	1.28E-06	1.47E-05	1.55	0.123	7	-1.35E-05	1.82E-04	-1.31	0.191		
	7	-1.20E-06	1.61E-05	-1.31	0.190	8	1.50E-05	1.32E-04	2.01	0.045	**	
	8	2.01E-07	1.39E-05	0.26	0.799	9	4.25E-06	9.37E-05	0.80	0.422		
	9	-1.44E-06	1.06E-05	-2.41	0.017	**	10	4.81E-06	7.01E-05	1.21	0.225	
10	9.59E-08	9.35E-06	0.18	0.856								
Depth	0	-1.28E-06	1.15E-06	-19.76	0.000	***	1	-9.45E-06	1.31E-05	-12.73	0.000	***
	1	-2.83E-07	3.68E-07	-13.58	0.000	***	2	-1.04E-06	1.16E-05	-1.58	0.114	
	2	-1.67E-07	2.93E-07	-10.09	0.000	***	3	-1.43E-06	6.06E-06	-4.18	0.000	***
	3	-8.75E-08	3.67E-07	-4.22	0.000	***	4	1.40E-07	6.08E-06	0.41	0.683	
	4	1.59E-08	7.32E-07	0.38	0.701	5	-1.49E-07	5.47E-06	-0.48	0.631		
	5	-7.91E-09	3.28E-07	-0.43	0.670	6	8.07E-07	6.98E-06	2.05	0.042	**	
	6	-3.15E-09	1.37E-07	-0.41	0.685	7	4.11E-07	3.41E-06	2.13	0.034	**	
	7	-8.60E-09	1.42E-07	-1.07	0.284	8	6.65E-07	3.39E-06	3.47	0.001	***	
	8	8.25E-09	2.23E-07	0.65	0.514	9	3.84E-07	3.38E-06	2.01	0.045	**	
	9	-1.24E-08	2.71E-07	-0.81	0.418	10	1.09E-06	4.86E-06	1.10	0.272		
10	2.29E-09	2.09E-07	0.19	0.846								



## A. SUPPLEMENT FOR CHAPTER 4

**Table A.5:** Adaptive Lasso Coefficients Shinhan Financials Group (055550 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.121	0.069	-31.22	0.000	***	1	15.395	5.227	52.11	0.000	***
	2	-0.061	0.054	-19.95	0.000	***	2	-17.290	5.566	-54.95	0.000	***
	3	-0.033	0.039	-14.73	0.000	***	3	-6.554	3.404	-34.07	0.000	***
	4	-0.017	0.026	-11.85	0.000	***	4	-4.850	2.678	-32.04	0.000	***
	5	-0.011	0.021	-9.09	0.000	***	5	-2.676	2.153	-21.99	0.000	***
	6	-0.008	0.018	-8.22	0.000	***	6	-1.537	1.724	-15.77	0.000	***
	7	-0.004	0.015	-4.49	0.000	***	7	-0.843	1.262	-11.83	0.000	***
	8	-0.003	0.015	-3.75	0.000	***	8	-0.514	0.955	-9.52	0.000	***
	9	0.000	0.012	-0.54	0.593		9	-0.173	0.519	-5.91	0.000	***
	10	-0.001	0.010	-1.20	0.231		10	-0.076	0.370	-3.61	0.000	***
Trades	0	1.10E-03	6.66E-04	29.10	0.000	***	1	0.284	0.069	72.89	0.000	***
	1	1.99E-04	2.01E-04	17.49	0.000	***	2	0.067	0.046	25.69	0.000	***
	2	1.71E-04	1.83E-04	16.53	0.000	***	3	0.069	0.014	87.00	0.000	***
	3	6.01E-05	8.62E-05	12.32	0.000	***	4	0.048	0.013	66.29	0.000	***
	4	3.26E-05	6.58E-05	8.76	0.000	***	5	0.039	0.013	54.14	0.000	***
	5	2.41E-05	6.10E-05	7.00	0.000	***	6	0.032	0.012	48.79	0.000	***
	6	1.10E-05	3.67E-05	5.28	0.000	***	7	0.029	0.011	44.38	0.000	***
	7	8.50E-06	3.64E-05	4.13	0.000	***	8	0.025	0.012	37.92	0.000	***
	8	3.91E-06	3.00E-05	2.31	0.022	**	9	0.023	0.012	34.36	0.000	***
	9	4.77E-06	2.45E-05	3.45	0.001	***	10	0.028	0.011	44.74	0.000	***
10	-6.39E-07	1.91E-05	-0.59	0.554								
Durations	0	4.39E-06	3.77E-06	20.59	0.000	***	1	3.73E-06	2.88E-04	0.23	0.819	
	1	-1.08E-07	6.68E-06	-0.29	0.776		2	8.24E-06	2.51E-04	0.58	0.562	
	2	5.54E-08	3.03E-06	0.32	0.747		3	7.07E-06	1.51E-04	0.83	0.408	
	3	6.63E-08	3.91E-06	0.30	0.764		4	-5.50E-06	8.19E-05	-1.19	0.236	
	4	1.48E-07	3.67E-06	0.71	0.476		5	-6.75E-07	6.34E-05	-0.19	0.851	
	5	4.35E-08	3.10E-06	0.25	0.804		6	-5.51E-07	5.50E-05	-0.18	0.859	
	6	7.73E-08	2.14E-06	0.64	0.522		7	3.09E-06	4.72E-05	1.16	0.248	
	7	-8.42E-08	2.18E-06	-0.68	0.494		8	3.93E-06	4.24E-05	1.64	0.102	
	8	-3.06E-07	2.90E-06	-1.87	0.063	*	9	1.12E-06	3.49E-05	0.57	0.570	
	9	-2.93E-08	2.86E-06	-0.18	0.857		10	1.67E-06	4.20E-05	0.71	0.481	
10	-1.17E-07	1.70E-06	-1.21	0.226								
Spreads	0	6.54E-05	1.57E-04	7.36	0.000	***	1	1.27E-03	4.63E-03	4.84	0.000	***
	1	4.12E-05	1.29E-04	5.62	0.000	***	2	1.83E-04	2.91E-03	1.11	0.266	
	2	1.22E-05	9.98E-05	2.17	0.031	**	3	2.55E-04	1.94E-03	2.33	0.021	**
	3	7.01E-06	6.56E-05	1.89	0.060	*	4	5.82E-05	1.18E-03	0.87	0.383	
	4	1.10E-06	6.01E-05	0.32	0.747		5	1.09E-07	7.39E-04	0.00	0.998	
	5	1.79E-06	4.58E-05	0.69	0.490		6	1.23E-04	9.94E-04	2.20	0.029	**
	6	6.89E-07	5.34E-05	0.23	0.820		7	-7.80E-06	5.20E-04	-0.27	0.791	
	7	-2.24E-06	5.50E-05	-0.72	0.471		8	6.65E-05	8.09E-04	1.45	0.147	
	8	2.32E-07	4.31E-05	0.10	0.924		9	-3.74E-05	6.86E-04	-0.97	0.335	
	9	1.30E-06	3.36E-05	0.69	0.494		10	-1.87E-06	5.65E-04	-0.06	0.953	
10	-1.71E-06	2.89E-05	-1.05	0.296								
Depth	0	-4.35E-07	3.22E-07	-23.91	0.000	***	1	-1.19E-05	1.15E-05	-18.40	0.000	***
	1	-7.80E-08	1.19E-07	-11.59	0.000	***	2	-7.57E-08	5.33E-06	-0.25	0.802	
	2	-2.59E-08	8.36E-08	-5.48	0.000	***	3	-3.02E-07	2.38E-06	-2.25	0.025	**
	3	-7.81E-09	5.18E-08	-2.66	0.008	***	4	4.04E-07	3.01E-06	2.37	0.018	**
	4	3.56E-10	3.39E-08	0.19	0.853		5	3.99E-07	2.33E-06	3.03	0.003	***
	5	-1.93E-11	3.61E-08	-0.01	0.992		6	6.92E-07	2.57E-06	4.77	0.000	***
	6	7.06E-11	6.24E-08	0.02	0.984		7	1.04E-06	3.60E-06	5.12	0.000	***
	7	8.79E-10	2.23E-08	0.70	0.486		8	7.88E-07	2.89E-06	4.82	0.000	***
	8	6.54E-09	4.13E-08	2.80	0.005	***	9	8.37E-07	2.68E-06	5.53	0.000	***
	9	7.51E-09	6.48E-08	2.05	0.041	**	10	1.17E-06	3.25E-06	1.10	0.272	
10	3.77E-09	6.52E-08	1.02	0.308								

**Table A.6:** Adaptive Lasso Coefficients Kia Motors (000270 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.101	0.063	-28.55	0.000	***	1	9.024	4.510	35.40	0.000	***
	2	-0.062	0.040	-27.58	0.000	***	2	-16.405	5.216	-55.64	0.000	***
	3	-0.040	0.035	-20.60	0.000	***	3	-9.548	4.488	-37.64	0.000	***
	4	-0.029	0.026	-19.46	0.000	***	4	-7.695	3.327	-40.92	0.000	***
	5	-0.016	0.025	-11.33	0.000	***	5	-5.637	3.056	-32.64	0.000	***
	6	-0.011	0.020	-9.83	0.000	***	6	-4.266	2.649	-28.49	0.000	***
	7	-0.007	0.014	-9.01	0.000	***	7	-3.027	2.239	-23.91	0.000	***
	8	-0.007	0.013	-9.17	0.000	***	8	-2.072	1.790	-20.48	0.000	***
	9	-0.005	0.015	-5.43	0.000	***	9	-1.251	1.301	-17.01	0.000	***
	10	-0.003	0.016	-3.08	0.002	***	10	-0.512	0.799	-11.35	0.000	***
Trades	0	8.83E-04	7.96E-04	19.64	0.000	***	1	0.253	0.104	42.84	0.000	***
	1	1.35E-04	1.67E-04	14.28	0.000	***	2	0.076	0.046	29.21	0.000	***
	2	1.48E-04	1.31E-04	19.99	0.000	***	3	0.085	0.017	87.16	0.000	***
	3	7.63E-05	1.04E-04	12.99	0.000	***	4	0.061	0.016	66.42	0.000	***
	4	5.07E-05	7.81E-05	11.49	0.000	***	5	0.051	0.016	57.71	0.000	***
	5	3.90E-05	6.47E-05	10.66	0.000	***	6	0.043	0.016	46.39	0.000	***
	6	2.35E-05	5.01E-05	8.30	0.000	***	7	0.039	0.015	46.25	0.000	***
	7	1.57E-05	4.71E-05	5.87	0.000	***	8	0.034	0.016	37.96	0.000	***
	8	1.66E-05	3.94E-05	7.48	0.000	***	9	0.033	0.015	38.14	0.000	***
	9	1.16E-05	4.16E-05	4.93	0.000	***	10	0.034	0.016	37.80	0.000	***
Durations	0	3.57E-06	3.51E-06	18.02	0.000	***	1	4.21E-06	4.68E-05	1.59	0.112	
	1	1.97E-07	2.82E-06	1.23	0.218		2	7.22E-06	3.98E-05	3.21	0.001	***
	2	1.60E-07	3.46E-06	0.82	0.414		3	2.63E-06	4.86E-05	0.96	0.340	
	3	-2.56E-07	4.28E-06	-1.06	0.291		4	9.48E-06	8.76E-05	1.92	0.056	*
	4	2.60E-07	3.56E-06	1.29	0.198		5	4.67E-08	1.22E-04	0.01	0.995	
	5	2.56E-07	3.72E-06	1.22	0.224		6	2.36E-06	6.08E-05	0.69	0.492	
	6	-2.43E-07	2.42E-06	-1.78	0.076	*	7	9.69E-06	7.21E-05	2.38	0.018	**
	7	-2.99E-08	2.45E-06	-0.22	0.830		8	7.54E-07	4.71E-05	0.28	0.777	
	8	-2.13E-07	2.14E-06	-1.76	0.080	*	9	3.61E-06	5.57E-05	1.15	0.253	
	9	5.86E-09	1.69E-06	0.06	0.951		10	-1.38E-06	4.40E-05	-0.56	0.579	
Spreads	0	1.38E-04	3.94E-04	6.20	0.000	***	1	1.68E-03	7.35E-03	4.04	0.000	***
	1	2.72E-05	3.94E-04	1.22	0.222		2	4.62E-04	3.29E-03	2.48	0.013	**
	2	4.46E-05	2.91E-04	2.71	0.007	***	3	4.72E-04	3.61E-03	2.31	0.021	**
	3	1.69E-05	2.32E-04	1.29	0.197		4	2.19E-04	3.62E-03	1.07	0.284	
	4	1.51E-05	2.06E-04	1.30	0.195		5	3.14E-04	2.45E-03	2.27	0.024	**
	5	3.39E-06	2.53E-04	0.24	0.813		6	2.44E-04	2.91E-03	1.48	0.139	
	6	3.00E-06	2.13E-04	0.25	0.804		7	3.57E-04	2.41E-03	2.62	0.009	***
	7	2.08E-06	1.75E-04	0.21	0.834		8	1.97E-04	1.37E-03	2.54	0.011	**
	8	3.83E-06	1.75E-04	0.39	0.698		9	5.47E-05	1.76E-03	0.55	0.582	
	9	-6.16E-06	1.94E-04	-0.56	0.575		10	1.63E-04	1.21E-03	2.39	0.018	**
Depth	0	-1.45E-07	1.50E-07	-17.12	0.000	***	1	-5.08E-06	4.63E-06	-19.44	0.000	***
	1	-2.50E-08	4.84E-08	-9.12	0.000	***	2	-1.42E-06	2.75E-06	-9.12	0.000	***
	2	-1.20E-08	2.86E-08	-7.42	0.000	***	3	-7.49E-07	1.76E-06	-7.52	0.000	***
	3	-8.33E-09	2.44E-08	-6.04	0.000	***	4	-2.22E-07	1.59E-06	-2.47	0.014	**
	4	-3.51E-09	2.51E-08	-2.47	0.014	**	5	-3.77E-08	8.56E-07	-0.78	0.436	
	5	-2.52E-09	2.32E-08	-1.92	0.055	*	6	1.02E-07	9.59E-07	1.89	0.060	*
	6	-1.44E-09	1.44E-08	-1.76	0.079	*	7	2.37E-07	1.35E-06	3.11	0.002	***
	7	1.75E-09	4.29E-08	0.72	0.472		8	3.67E-07	1.43E-06	4.55	0.000	***
	8	-2.41E-09	4.76E-08	-0.90	0.370		9	5.05E-07	1.61E-06	5.53	0.000	***
	9	2.64E-10	1.24E-08	0.38	0.706		10	7.19E-07	1.70E-06	1.10	0.272	

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.7:** Adaptive Lasso Coefficients SK Hynix (000660 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.075	0.058	-22.72	0.000	***	1	9.086	4.779	33.63	0.000	***
	2	-0.049	0.039	-21.94	0.000	***	2	-19.283	6.994	-48.78	0.000	***
	3	-0.031	0.031	-17.82	0.000	***	3	-11.709	4.944	-41.90	0.000	***
	4	-0.024	0.032	-13.30	0.000	***	4	-9.977	4.092	-43.13	0.000	***
	5	-0.016	0.024	-11.35	0.000	***	5	-7.652	3.473	-38.98	0.000	***
	6	-0.011	0.019	-9.80	0.000	***	6	-5.963	2.794	-37.75	0.000	***
	7	-0.010	0.022	-7.65	0.000	***	7	-4.443	2.311	-34.02	0.000	***
	8	-0.007	0.020	-5.79	0.000	***	8	-3.110	1.861	-29.57	0.000	***
	9	-0.005	0.017	-5.48	0.000	***	9	-2.041	1.515	-23.84	0.000	***
	10	-0.002	0.017	-2.28	0.023	**	10	-0.998	1.120	-15.76	0.000	***
Trades	0	5.15E-04	4.68E-04	19.47	0.000	***	1	0.231	0.080	50.94	0.000	***
	1	9.91E-05	9.45E-05	18.56	0.000	***	2	0.072	0.053	24.01	0.000	***
	2	1.03E-04	9.12E-05	19.96	0.000	***	3	0.085	0.019	78.45	0.000	***
	3	5.63E-05	5.79E-05	17.20	0.000	***	4	0.067	0.017	69.80	0.000	***
	4	3.84E-05	5.90E-05	11.51	0.000	***	5	0.058	0.015	68.53	0.000	***
	5	2.78E-05	4.38E-05	11.26	0.000	***	6	0.051	0.014	64.67	0.000	***
	6	1.90E-05	3.65E-05	9.18	0.000	***	7	0.045	0.014	59.10	0.000	***
	7	1.87E-05	3.82E-05	8.68	0.000	***	8	0.043	0.013	56.45	0.000	***
	8	1.58E-05	3.59E-05	7.77	0.000	***	9	0.038	0.014	48.36	0.000	***
	9	1.07E-05	2.93E-05	6.45	0.000	***	10	0.041	0.013	57.92	0.000	***
10	5.48E-06	1.87E-05	5.17	0.000	***							
Durations	0	2.82E-06	2.69E-06	18.51	0.000	***	1	-2.18E-06	3.10E-05	-1.25	0.214	
	1	3.99E-07	3.07E-06	2.30	0.022	**	2	-1.61E-06	4.16E-05	-0.68	0.495	
	2	1.61E-07	2.69E-06	1.06	0.292		3	3.31E-06	7.07E-05	0.83	0.408	
	3	8.78E-08	2.88E-06	0.54	0.590		4	5.31E-06	7.37E-05	1.28	0.203	
	4	8.40E-08	2.69E-06	0.55	0.581		5	5.80E-06	6.61E-05	1.55	0.122	
	5	6.52E-08	2.72E-06	0.42	0.672		6	6.55E-06	6.66E-05	1.74	0.083	*
	6	-2.80E-07	3.70E-06	-1.34	0.183		7	6.94E-06	6.34E-05	1.94	0.054	*
	7	1.95E-07	2.19E-06	1.58	0.115		8	4.21E-07	6.01E-05	0.12	0.901	
	8	-3.53E-08	2.20E-06	-0.28	0.777		9	9.27E-06	5.65E-05	2.90	0.004	***
	9	-1.56E-07	4.06E-06	-0.68	0.498		10	2.99E-06	6.15E-05	-0.86	0.390	
10	8.26E-08	1.53E-06	0.95	0.342								
Spreads	0	7.67E-05	4.08E-04	3.33	0.001	***	1	4.00E-04	1.17E-02	0.61	0.545	
	1	6.08E-05	2.55E-04	4.22	0.000	***	2	-9.45E-05	6.88E-03	-0.24	0.808	
	2	4.56E-05	2.81E-04	2.87	0.004	***	3	-5.95E-05	4.04E-03	-0.26	0.795	
	3	3.72E-05	2.27E-04	2.90	0.004	***	4	4.28E-04	3.82E-03	1.98	0.048	**
	4	4.53E-05	2.42E-04	3.31	0.001	***	5	2.60E-04	3.33E-03	1.38	0.168	
	5	2.49E-05	2.74E-04	1.60	0.110		6	4.97E-04	3.96E-03	2.22	0.027	**
	6	1.78E-05	2.68E-04	1.18	0.240		7	8.28E-04	5.76E-03	2.55	0.011	**
	7	1.29E-05	2.52E-04	0.91	0.366		8	7.28E-04	3.09E-03	4.17	0.000	***
	8	7.52E-06	2.48E-04	0.54	0.592		9	8.92E-04	4.34E-03	3.64	0.000	***
	9	4.56E-05	2.37E-04	3.40	0.001	***	10	6.73E-04	3.67E-03	3.24	0.001	***
10	7.80E-06	1.85E-04	0.75	0.456								
Depth	0	-6.78E-08	8.02E-08	-14.95	0.000	***	1	-4.71E-06	3.68E-06	-22.67	0.000	***
	1	-8.93E-09	1.34E-08	-11.79	0.000	***	2	-1.02E-06	1.43E-06	-12.62	0.000	***
	2	-5.74E-09	1.39E-08	-7.28	0.000	***	3	-4.17E-07	9.60E-07	-7.68	0.000	***
	3	-2.91E-09	9.94E-09	-5.18	0.000	***	4	-9.59E-08	7.24E-07	-2.34	0.020	**
	4	-2.25E-09	7.47E-09	-5.33	0.000	***	5	5.81E-08	6.02E-07	1.71	0.089	*
	5	-7.49E-10	4.78E-09	-2.78	0.006	***	6	1.85E-07	7.82E-07	4.19	0.000	***
	6	-2.23E-10	4.53E-09	-0.87	0.385		7	2.59E-07	9.10E-07	5.04	0.000	***
	7	-6.52E-10	6.94E-09	-1.66	0.097	*	8	4.12E-07	1.03E-06	7.08	0.000	***
	8	-5.09E-10	5.32E-09	-1.69	0.091	*	9	5.07E-07	1.12E-06	7.98	0.000	***
	9	3.22E-10	7.77E-09	0.73	0.464		10	9.38E-07	1.24E-06	1.10	0.272	
10	1.15E-09	4.88E-09	4.16	0.000	***							

**Table A.8:** Adaptive Lasso Coefficients Hyundai Heavy Industries (009540 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.115	0.058	-35.48	0.000	***	1	8.254	2.676	54.57	0.000	***
	2	-0.066	0.040	-29.12	0.000	***	2	-8.474	2.653	-56.50	0.000	***
	3	-0.043	0.029	-25.93	0.000	***	3	-4.121	1.637	-44.54	0.000	***
	4	-0.027	0.027	-17.68	0.000	***	4	-3.329	1.406	-41.88	0.000	***
	5	-0.015	0.019	-14.46	0.000	***	5	-2.188	1.127	-34.35	0.000	***
	6	-0.012	0.020	-10.36	0.000	***	6	-1.401	1.009	-24.56	0.000	***
	7	-0.007	0.016	-7.85	0.000	***	7	-0.874	0.811	-19.07	0.000	***
	8	-0.006	0.013	-7.44	0.000	***	8	-0.506	0.614	-14.57	0.000	***
	9	-0.003	0.012	-4.45	0.000	***	9	-0.219	0.395	-9.80	0.000	***
	10	-0.002	0.010	-3.98	0.000	***	10	-0.092	0.237	-6.91	0.000	***
Trades	0	1.55E-03	9.60E-04	28.61	0.000	***	1	0.237	0.077	54.68	0.000	***
	1	3.83E-04	2.80E-04	24.25	0.000	***	2	0.068	0.042	28.76	0.000	***
	2	3.37E-04	2.38E-04	25.05	0.000	***	3	0.070	0.012	106.67	0.000	***
	3	1.85E-04	1.73E-04	18.95	0.000	***	4	0.052	0.012	76.52	0.000	***
	4	1.13E-04	1.34E-04	14.85	0.000	***	5	0.043	0.011	66.59	0.000	***
	5	8.73E-05	1.41E-04	10.97	0.000	***	6	0.035	0.011	55.59	0.000	***
	6	4.16E-05	7.59E-05	9.70	0.000	***	7	0.031	0.011	49.32	0.000	***
	7	4.10E-05	9.58E-05	7.58	0.000	***	8	0.028	0.011	45.06	0.000	***
	8	2.02E-05	5.74E-05	6.22	0.000	***	9	0.025	0.011	39.05	0.000	***
	9	2.18E-05	6.94E-05	5.55	0.000	***	10	0.028	0.010	50.14	0.000	***
Durations	0	4.57E-06	3.91E-06	20.69	0.000	***	1	9.15E-06	9.38E-05	1.72	0.086	*
	1	2.03E-07	4.38E-06	0.82	0.412		2	1.02E-05	7.59E-05	2.37	0.018	**
	2	1.84E-07	4.97E-06	0.65	0.514		3	2.63E-06	5.73E-05	0.81	0.417	
	3	-7.11E-08	4.03E-06	-0.31	0.755		4	3.46E-06	5.09E-05	1.21	0.229	
	4	1.64E-07	2.27E-06	1.28	0.200		5	5.10E-08	3.79E-05	0.02	0.981	
	5	7.68E-08	2.32E-06	0.58	0.559		6	7.64E-07	3.18E-05	0.42	0.672	
	6	-1.40E-08	2.23E-06	-0.11	0.911		7	1.72E-08	2.44E-05	0.01	0.990	
	7	-3.88E-08	2.71E-06	-0.25	0.800		8	6.80E-08	2.81E-05	0.04	0.966	
	8	1.19E-07	1.75E-06	1.20	0.230		9	1.29E-06	2.36E-05	0.96	0.336	
	9	1.59E-08	8.19E-07	0.34	0.732		10	8.48E-07	1.51E-05	0.99	0.323	
Spreads	0	7.93E-06	3.23E-05	4.35	0.000	***	1	3.64E-05	3.72E-04	1.73	0.084	*
	1	3.46E-06	1.37E-05	4.47	0.000	***	2	3.76E-06	1.92E-04	0.35	0.729	
	2	2.26E-06	1.12E-05	3.55	0.000	***	3	-8.08E-06	1.37E-04	-1.04	0.299	
	3	9.80E-07	1.44E-05	1.20	0.230		4	4.13E-06	8.86E-05	0.82	0.410	
	4	7.49E-07	9.37E-06	1.41	0.158		5	-1.61E-06	1.15E-04	-0.25	0.804	
	5	1.16E-06	8.97E-06	2.28	0.023	**	6	1.75E-06	7.44E-05	0.42	0.678	
	6	5.25E-07	8.85E-06	1.05	0.295		7	7.52E-06	8.10E-05	1.64	0.101	
	7	1.02E-06	1.03E-05	1.75	0.081	*	8	2.61E-06	7.78E-05	0.59	0.553	
	8	4.24E-08	1.08E-05	0.07	0.945		9	1.14E-05	1.04E-04	1.94	0.054	*
	9	7.41E-07	9.44E-06	1.39	0.166		10	5.35E-06	4.76E-05	1.99	0.048	**
Depth	0	-1.54E-06	1.25E-06	-21.79	0.000	***	1	-1.39E-05	1.71E-05	-14.42	0.000	***
	1	-3.32E-07	4.07E-07	-14.45	0.000	***	2	-2.31E-06	9.24E-06	-4.42	0.000	***
	2	-1.67E-07	2.82E-07	-10.45	0.000	***	3	-1.32E-06	5.55E-06	-4.22	0.000	***
	3	-9.34E-08	1.86E-07	-8.87	0.000	***	4	-4.33E-07	5.37E-06	-1.43	0.155	
	4	-4.31E-08	1.39E-07	-5.50	0.000	***	5	4.97E-07	6.91E-06	1.27	0.204	
	5	-6.26E-09	1.92E-07	-0.58	0.565		6	1.21E-06	6.55E-06	3.26	0.001	***
	6	-2.41E-08	1.74E-07	-2.45	0.015	**	7	1.03E-06	5.55E-06	3.29	0.001	***
	7	-4.37E-10	8.31E-08	-0.09	0.926		8	1.58E-06	6.82E-06	4.11	0.000	***
	8	-4.78E-09	1.37E-07	-0.62	0.537		9	1.92E-06	6.55E-06	5.19	0.000	***
	9	-1.74E-08	1.24E-07	-2.48	0.014	**	10	2.72E-06	7.37E-06	1.10	0.272	
10	1.34E-08	2.43E-07	0.98	0.329								

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.9:** Adaptive Lasso Coefficients KEPCO (015760 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.131	0.054	-42.86	0.000	***	1	15.408	4.793	56.87	0.000	***
	2	-0.060	0.035	-30.55	0.000	***	2	-19.608	7.013	-49.46	0.000	***
	3	-0.039	0.043	-16.37	0.000	***	3	-7.718	4.160	-32.82	0.000	***
	4	-0.019	0.024	-13.85	0.000	***	4	-5.532	3.404	-28.76	0.000	***
	5	-0.012	0.025	-8.89	0.000	***	5	-3.142	2.575	-21.59	0.000	***
	6	-0.009	0.020	-8.01	0.000	***	6	-1.883	1.952	-17.06	0.000	***
	7	-0.005	0.013	-6.37	0.000	***	7	-0.990	1.376	-12.73	0.000	***
	8	-0.003	0.011	-5.28	0.000	***	8	-0.541	0.945	-10.13	0.000	***
	9	-0.003	0.011	-4.64	0.000	***	9	-0.167	0.498	-5.92	0.000	***
	10	-0.002	0.009	-3.83	0.000	***	10	-0.081	0.374	-3.81	0.000	***
Trades	0	1.04E-03	7.06E-04	26.06	0.000	***	1	0.294	0.072	72.75	0.000	***
	1	1.52E-04	1.95E-04	13.77	0.000	***	2	0.071	0.048	25.79	0.000	***
	2	1.44E-04	1.70E-04	14.99	0.000	***	3	0.072	0.016	79.00	0.000	***
	3	3.93E-05	7.01E-05	9.92	0.000	***	4	0.047	0.015	54.99	0.000	***
	4	2.21E-05	7.57E-05	5.18	0.000	***	5	0.039	0.013	53.72	0.000	***
	5	1.52E-05	5.22E-05	5.15	0.000	***	6	0.030	0.014	38.80	0.000	***
	6	1.11E-05	3.98E-05	4.95	0.000	***	7	0.026	0.013	35.59	0.000	***
	7	6.15E-06	2.62E-05	4.16	0.000	***	8	0.024	0.012	34.68	0.000	***
	8	2.95E-06	1.45E-05	3.60	0.000	***	9	0.020	0.013	26.89	0.000	***
	9	2.83E-06	3.32E-05	1.51	0.132		10	0.026	0.013	35.65	0.000	***
Durations	0	3.64E-06	3.03E-06	21.25	0.000	***	1	-5.79E-06	6.17E-05	-1.66	0.098	*
	1	-5.66E-08	2.66E-06	-0.38	0.707		2	7.86E-06	7.07E-05	1.97	0.050	*
	2	-2.04E-07	3.18E-06	-1.14	0.257		3	8.21E-07	6.72E-05	0.22	0.829	
	3	-2.36E-07	2.85E-06	-1.46	0.144		4	3.05E-06	5.77E-05	0.93	0.351	
	4	-1.26E-07	2.94E-06	-0.76	0.450		5	-1.70E-06	4.65E-05	-0.65	0.517	
	5	6.53E-08	2.69E-06	0.43	0.668		6	-3.49E-06	6.64E-05	-0.93	0.354	
	6	-6.08E-08	1.71E-06	-0.63	0.529		7	2.54E-06	4.00E-05	1.12	0.262	
	7	-6.89E-08	2.45E-06	-0.50	0.620		8	-4.43E-07	2.64E-05	-0.30	0.767	
	8	-2.16E-07	1.68E-06	-2.28	0.023	**	9	1.08E-06	3.89E-05	0.49	0.623	
	9	-1.24E-08	1.34E-06	-0.16	0.870		10	-4.20E-06	3.45E-05	-2.15	0.032	**
Spreads	0	9.32E-05	1.57E-04	10.52	0.000	***	1	1.14E-03	5.49E-03	3.67	0.000	***
	1	3.47E-05	1.62E-04	3.78	0.000	***	2	3.56E-04	3.53E-03	1.79	0.075	*
	2	1.75E-05	1.24E-04	2.49	0.013	**	3	2.30E-04	2.06E-03	1.98	0.049	**
	3	1.34E-05	9.37E-05	2.53	0.012	**	4	9.49E-05	1.32E-03	1.27	0.204	
	4	1.19E-05	9.51E-05	2.22	0.027	**	5	1.19E-04	1.51E-03	1.38	0.167	
	5	-2.35E-06	1.07E-04	-0.39	0.698		6	1.70E-05	1.39E-03	0.22	0.829	
	6	7.25E-06	7.82E-05	1.64	0.102		7	-4.95E-05	1.59E-03	-0.55	0.582	
	7	2.82E-06	6.66E-05	0.75	0.454		8	8.33E-05	1.09E-03	1.36	0.175	
	8	2.39E-06	7.49E-05	0.57	0.572		9	-1.08E-06	7.93E-04	-0.02	0.981	
	9	-5.77E-06	5.37E-05	-1.90	0.058	*	10	6.07E-05	1.03E-03	1.04	0.299	
Depth	0	-3.61E-07	3.08E-07	-20.71	0.000	***	1	-8.47E-06	8.08E-06	-18.53	0.000	***
	1	-4.98E-08	9.26E-08	-9.51	0.000	***	2	-6.40E-07	2.66E-06	-4.26	0.000	***
	2	-2.28E-08	5.63E-08	-7.15	0.000	***	3	-4.98E-07	2.16E-06	-4.07	0.000	***
	3	-1.58E-09	5.13E-08	-0.55	0.585		4	-4.26E-08	1.62E-06	-0.47	0.642	
	4	-8.04E-10	4.02E-08	-0.35	0.724		5	1.44E-07	1.96E-06	1.30	0.196	
	5	-5.19E-10	2.72E-08	-0.34	0.736		6	2.77E-07	1.80E-06	2.72	0.007	***
	6	3.88E-09	3.70E-08	1.85	0.065	*	7	2.45E-07	1.66E-06	2.61	0.009	***
	7	-1.59E-09	4.15E-08	-0.68	0.499		8	3.65E-07	1.50E-06	4.31	0.000	***
	8	2.86E-09	1.94E-08	2.61	0.010	***	9	3.72E-07	1.93E-06	3.41	0.001	***
	9	4.51E-09	3.97E-08	2.01	0.045	**	10	5.34E-07	2.10E-06	1.10	0.272	
10	3.14E-09	3.32E-08	1.67	0.095	*							

**Table A.10:** Adaptive Lasso Coefficients SK Telecom (017670 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.120	0.041	-51.38	0.000	***	1	7.429	1.592	82.58	0.000	***
	2	-0.063	0.027	-41.28	0.000	***	2	-7.076	2.910	-43.02	0.000	***
	3	-0.039	0.024	-29.47	0.000	***	3	-2.622	1.568	-29.59	0.000	***
	4	-0.026	0.020	-23.29	0.000	***	4	-1.828	1.349	-23.99	0.000	***
	5	-0.017	0.018	-16.27	0.000	***	5	-0.992	1.034	-16.97	0.000	***
	6	-0.012	0.018	-11.96	0.000	***	6	-0.571	0.758	-13.32	0.000	***
	7	-0.007	0.013	-9.83	0.000	***	7	-0.355	0.594	-10.57	0.000	***
	8	-0.007	0.013	-9.21	0.000	***	8	-0.195	0.448	-7.71	0.000	***
	9	-0.004	0.010	-7.07	0.000	***	9	-0.096	0.291	-5.86	0.000	***
	10	-0.003	0.010	-5.64	0.000	***	10	-0.069	0.232	-5.29	0.000	***
Trades	0	1.52E-03	1.18E-03	22.74	0.000	***	1	0.274	0.076	63.89	0.000	***
	1	1.94E-04	2.51E-04	13.63	0.000	***	2	0.078	0.037	37.39	0.000	***
	2	1.59E-04	2.19E-04	12.88	0.000	***	3	0.066	0.018	64.00	0.000	***
	3	8.14E-05	1.44E-04	10.01	0.000	***	4	0.045	0.016	51.03	0.000	***
	4	3.82E-05	9.19E-05	7.35	0.000	***	5	0.035	0.015	41.59	0.000	***
	5	2.05E-05	6.76E-05	5.36	0.000	***	6	0.028	0.015	32.70	0.000	***
	6	2.01E-05	6.29E-05	5.64	0.000	***	7	0.025	0.014	31.51	0.000	***
	7	1.60E-05	5.53E-05	5.11	0.000	***	8	0.022	0.016	25.26	0.000	***
	8	1.03E-05	4.86E-05	3.76	0.000	***	9	0.021	0.014	25.80	0.000	***
	9	1.04E-05	4.63E-05	3.98	0.000	***	10	0.025	0.014	31.83	0.000	***
Durations	0	3.91E-06	4.20E-06	16.47	0.000	***	1	1.62E-06	3.82E-05	0.75	0.452	
	1	-1.07E-07	1.71E-06	-1.11	0.266		2	-2.68E-06	2.93E-05	-1.62	0.106	
	2	-1.67E-07	1.63E-06	-1.81	0.071	*	3	-1.19E-07	1.62E-05	-0.13	0.897	
	3	-7.32E-08	2.69E-06	-0.48	0.630		4	-1.84E-06	1.82E-05	-1.79	0.074	*
	4	4.09E-08	1.21E-06	0.60	0.549		5	-6.08E-07	1.30E-05	-0.83	0.407	
	5	-6.91E-08	1.64E-06	-0.75	0.455		6	1.01E-06	1.32E-05	1.35	0.177	
	6	-4.52E-08	9.14E-07	-0.88	0.382		7	9.31E-08	6.06E-06	0.27	0.786	
	7	-1.61E-08	1.44E-06	-0.20	0.844		8	2.49E-08	4.52E-06	0.10	0.922	
	8	-2.49E-08	8.81E-07	-0.50	0.617		9	5.21E-07	8.81E-06	1.05	0.296	
	9	-3.41E-10	9.88E-07	-0.01	0.995		10	-5.42E-07	9.49E-06	-1.01	0.314	
Spreads	0	7.05E-06	2.15E-05	5.80	0.000	***	1	4.77E-05	2.15E-04	3.92	0.000	***
	1	6.61E-06	2.02E-05	5.77	0.000	***	2	-7.24E-07	1.26E-04	-0.10	0.919	
	2	1.45E-06	1.71E-05	1.50	0.135		3	1.31E-05	1.11E-04	2.10	0.037	**
	3	1.69E-06	2.09E-05	1.42	0.156		4	4.08E-06	1.03E-04	0.70	0.483	
	4	-3.48E-07	1.79E-05	-0.34	0.732		5	1.14E-06	8.08E-05	0.25	0.804	
	5	1.48E-06	1.48E-05	1.76	0.079	*	6	-2.56E-05	3.18E-04	-1.42	0.156	
	6	8.87E-07	1.09E-05	1.44	0.150		7	-6.83E-06	1.34E-04	-0.90	0.367	
	7	3.56E-07	1.12E-05	0.56	0.575		8	-2.48E-06	8.03E-05	-0.55	0.586	
	8	-5.41E-07	1.33E-05	-0.72	0.472		9	-9.47E-06	1.30E-04	-1.29	0.198	
	9	5.55E-07	9.75E-06	1.01	0.315		10	-3.14E-06	6.71E-05	-0.83	0.409	
Depth	0	-1.36E-06	1.48E-06	-16.25	0.000	***	1	-6.96E-06	1.38E-05	-8.94	0.000	***
	1	-1.54E-07	2.84E-07	-9.61	0.000	***	2	-9.51E-07	7.18E-06	-2.34	0.020	**
	2	-7.83E-08	1.78E-07	-7.77	0.000	***	3	-7.44E-07	5.39E-06	-2.44	0.015	**
	3	-5.88E-08	2.70E-07	-3.85	0.000	***	4	-5.03E-07	4.07E-06	-2.19	0.029	**
	4	-1.19E-08	1.13E-07	-1.86	0.064	*	5	-4.71E-07	4.78E-06	-1.74	0.083	*
	5	-9.94E-09	5.59E-08	-3.15	0.002	***	6	-2.30E-07	1.50E-06	-2.71	0.007	***
	6	-1.70E-08	1.12E-07	-2.68	0.008	***	7	8.74E-08	3.24E-06	0.48	0.634	
	7	-7.60E-09	9.12E-08	-1.47	0.141		8	2.54E-07	3.20E-06	1.41	0.161	
	8	-1.36E-08	1.70E-07	-1.41	0.160		9	9.61E-08	1.92E-06	0.88	0.378	
	9	-1.46E-08	1.55E-07	-1.66	0.098	*	10	4.56E-07	3.92E-06	1.10	0.272	
10	9.96E-10	1.02E-07	0.17	0.863								

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.11:** Adaptive Lasso Coefficients KT Corporation(030200 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.151	0.061	-44.17	0.000	***	1	18.467	5.691	57.41	0.000	***
	2	-0.072	0.047	-26.73	0.000	***	2	-21.637	7.456	-51.34	0.000	***
	3	-0.044	0.032	-24.44	0.000	***	3	-7.985	4.361	-32.40	0.000	***
	4	-0.025	0.028	-16.05	0.000	***	4	-5.733	3.159	-32.11	0.000	***
	5	-0.016	0.026	-11.08	0.000	***	5	-3.003	2.493	-21.31	0.000	***
	6	-0.010	0.017	-10.27	0.000	***	6	-1.776	1.925	-16.33	0.000	***
	7	-0.006	0.011	-9.75	0.000	***	7	-0.972	1.396	-12.31	0.000	***
	8	-0.004	0.011	-7.07	0.000	***	8	-0.505	0.951	-9.39	0.000	***
	9	-0.003	0.014	-3.62	0.000	***	9	-0.247	0.690	-6.34	0.000	***
	10	-0.002	0.011	-2.99	0.003	***	10	-0.103	0.422	-4.30	0.000	***
Trades	0	1.04E-03	7.90E-04	23.40	0.000	***	1	0.299	0.086	61.69	0.000	***
	1	1.43E-04	1.42E-04	17.85	0.000	***	2	0.069	0.047	25.92	0.000	***
	2	1.28E-04	1.60E-04	14.13	0.000	***	3	0.070	0.017	72.38	0.000	***
	3	4.28E-05	7.13E-05	10.60	0.000	***	4	0.044	0.017	46.68	0.000	***
	4	2.63E-05	5.78E-05	8.04	0.000	***	5	0.035	0.013	46.33	0.000	***
	5	1.38E-05	3.93E-05	6.20	0.000	***	6	0.027	0.015	31.81	0.000	***
	6	7.65E-06	2.84E-05	4.76	0.000	***	7	0.025	0.013	32.44	0.000	***
	7	2.28E-06	2.60E-05	1.55	0.122		8	0.020	0.013	28.12	0.000	***
	8	3.12E-06	1.92E-05	2.87	0.004	***	9	0.020	0.013	25.99	0.000	***
	9	2.38E-06	1.78E-05	2.36	0.019	**	10	0.024	0.015	29.12	0.000	***
10	5.67E-07	6.26E-06	1.60	0.110								
Durations	0	3.58E-06	3.53E-06	17.90	0.000	***	1	-3.00E-06	7.37E-05	-0.72	0.473	
	1	3.48E-09	3.11E-06	0.02	0.984		2	-6.16E-07	7.12E-05	-0.15	0.879	
	2	5.15E-08	2.82E-06	0.32	0.747		3	1.99E-06	5.74E-05	0.61	0.541	
	3	-5.64E-08	2.42E-06	-0.41	0.680		4	5.99E-06	6.23E-05	1.70	0.090	*
	4	-1.56E-07	2.76E-06	-1.00	0.317		5	1.32E-06	4.93E-05	0.47	0.637	
	5	-2.35E-07	1.91E-06	-2.17	0.030	**	6	-1.15E-06	4.40E-05	-0.46	0.644	
	6	1.15E-08	1.58E-06	0.13	0.898		7	-2.16E-06	3.32E-05	-1.15	0.252	
	7	-1.14E-07	1.28E-06	-1.57	0.118		8	3.83E-07	1.82E-05	0.37	0.710	
	8	-7.99E-08	1.36E-06	-1.04	0.300		9	-2.37E-06	2.49E-05	-1.68	0.094	*
	9	9.75E-08	1.12E-06	1.54	0.124		10	-3.06E-06	3.02E-05	-1.79	0.074	*
10	5.85E-08	1.55E-06	0.67	0.506								
Spreads	0	6.06E-05	1.27E-04	8.47	0.000	***	1	1.13E-03	5.47E-03	3.65	0.000	***
	1	2.42E-05	7.97E-05	5.38	0.000	***	2	3.37E-04	2.04E-03	2.92	0.004	***
	2	9.48E-06	6.33E-05	2.65	0.008	***	3	2.15E-04	1.26E-03	3.03	0.003	***
	3	1.59E-05	9.46E-05	2.97	0.003	***	4	2.77E-04	1.87E-03	2.63	0.009	***
	4	-2.53E-07	5.81E-05	-0.08	0.939		5	1.60E-04	1.71E-03	1.66	0.099	*
	5	1.50E-06	5.19E-05	0.51	0.608		6	9.45E-05	9.19E-04	1.82	0.070	*
	6	-2.73E-06	5.62E-05	-0.86	0.390		7	-1.69E-05	9.24E-04	-0.32	0.746	
	7	-8.46E-07	4.96E-05	-0.30	0.763		8	3.26E-06	6.47E-04	0.09	0.929	
	8	7.58E-06	7.12E-05	1.88	0.061	*	9	7.58E-06	7.31E-04	0.18	0.854	
	9	1.96E-06	4.55E-05	0.76	0.447		10	7.61E-05	6.20E-04	2.17	0.031	**
10	-3.34E-06	3.51E-05	-1.69	0.093	*							
Depth	0	-5.97E-07	5.82E-07	-18.16	0.000	***	1	-9.53E-06	1.01E-05	-16.65	0.000	***
	1	-6.05E-08	9.65E-08	-11.10	0.000	***	2	-4.85E-07	3.98E-06	-2.15	0.032	**
	2	-1.92E-08	6.87E-08	-4.95	0.000	***	3	-4.82E-08	2.93E-06	-0.29	0.771	
	3	-6.26E-09	3.69E-08	-3.00	0.003	***	4	-4.36E-08	2.37E-06	-0.33	0.745	
	4	-1.28E-09	4.01E-08	-0.56	0.573		5	1.91E-07	2.03E-06	1.66	0.098	*
	5	-8.16E-10	2.51E-08	-0.57	0.566		6	5.73E-07	4.02E-06	2.52	0.012	**
	6	2.73E-09	3.02E-08	1.60	0.111		7	5.49E-07	2.51E-06	3.86	0.000	***
	7	5.21E-10	5.00E-08	0.18	0.854		8	4.36E-07	1.98E-06	3.89	0.000	***
	8	3.21E-09	3.24E-08	1.75	0.081	*	9	6.60E-07	2.23E-06	5.24	0.000	***
	9	3.87E-09	4.79E-08	1.43	0.154		10	1.28E-06	3.67E-06	1.10	0.272	
10	8.42E-09	4.42E-08	3.37	0.001	***							

**Table A.12:** Adaptive Lasso Coefficients KT&G Corporation (033780 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.128	0.059	-38.56	0.000	***	1	13.129	3.664	63.40	0.000	***
	2	-0.055	0.034	-28.40	0.000	***	2	-12.937	4.698	-48.71	0.000	***
	3	-0.031	0.026	-21.10	0.000	***	3	-4.160	3.045	-24.17	0.000	***
	4	-0.017	0.026	-11.58	0.000	***	4	-2.827	2.323	-21.53	0.000	***
	5	-0.008	0.023	-6.38	0.000	***	5	-1.258	1.753	-12.70	0.000	***
	6	-0.005	0.014	-6.55	0.000	***	6	-0.685	1.278	-9.48	0.000	***
	7	-0.003	0.012	-4.16	0.000	***	7	-0.451	1.048	-7.61	0.000	***
	8	-0.002	0.019	-1.79	0.074	*	8	-0.193	0.632	-5.40	0.000	***
	9	-0.002	0.010	-2.79	0.006	***	9	-0.082	0.431	-3.38	0.001	***
	10	-0.002	0.012	-2.97	0.003	***	10	-0.041	0.285	-2.54	0.012	**
Trades	0	1.56E-03	9.40E-04	29.28	0.000	***	1	0.280	0.089	55.79	0.000	***
	1	3.21E-04	2.61E-04	21.81	0.000	***	2	0.067	0.046	25.78	0.000	***
	2	2.39E-04	1.90E-04	22.27	0.000	***	3	0.067	0.016	75.68	0.000	***
	3	1.06E-04	1.30E-04	14.46	0.000	***	4	0.041	0.016	45.66	0.000	***
	4	5.66E-05	1.11E-04	9.01	0.000	***	5	0.033	0.015	38.12	0.000	***
	5	2.79E-05	7.13E-05	6.92	0.000	***	6	0.026	0.015	31.29	0.000	***
	6	1.28E-05	6.43E-05	3.52	0.000	***	7	0.023	0.013	30.32	0.000	***
	7	1.55E-05	9.91E-05	2.77	0.006	***	8	0.019	0.013	25.28	0.000	***
	8	5.82E-06	4.10E-05	2.51	0.013	**	9	0.020	0.014	24.88	0.000	***
	9	5.96E-06	4.48E-05	2.35	0.019	**	10	0.021	0.014	27.32	0.000	***
Durations	0	5.04E-06	4.22E-06	21.10	0.000	***	1	-1.14E-06	8.06E-05	-0.25	0.803	
	1	2.57E-09	2.12E-06	0.02	0.983	2	4.31E-06	5.46E-05	1.40	0.164		
	2	3.79E-08	2.03E-06	0.33	0.741	3	-1.34E-06	3.69E-05	-0.64	0.520		
	3	-1.55E-07	3.53E-06	-0.78	0.439	4	7.00E-08	3.58E-05	0.03	0.972		
	4	5.18E-08	1.80E-06	0.51	0.611	5	-3.55E-06	3.95E-05	-1.59	0.113		
	5	-4.02E-08	9.16E-07	-0.78	0.438	6	4.86E-07	1.67E-05	0.52	0.606		
	6	3.80E-09	1.12E-06	0.06	0.952	7	-7.25E-07	1.06E-05	-1.21	0.228		
	7	-4.31E-08	1.12E-06	-0.68	0.495	8	7.33E-07	1.19E-05	1.09	0.275		
	8	-1.31E-07	1.89E-06	-1.22	0.222	9	1.03E-07	2.03E-05	0.09	0.929		
	9	-6.89E-08	1.11E-06	-1.10	0.273	10	-9.67E-07	2.74E-05	-0.63	0.532		
Spreads	0	3.51E-05	5.46E-05	11.36	0.000	***	1	3.90E-04	1.05E-03	6.57	0.000	***
	1	1.21E-05	4.73E-05	4.53	0.000	***	2	6.98E-05	6.74E-04	1.83	0.068	*
	2	6.00E-06	4.65E-05	2.28	0.023	**	3	2.41E-05	6.00E-04	0.71	0.478	
	3	4.04E-06	4.29E-05	1.67	0.097	*	4	2.41E-05	5.24E-04	0.81	0.417	
	4	-4.81E-07	2.59E-05	-0.33	0.742	5	-5.57E-06	4.55E-04	-0.22	0.829		
	5	-1.95E-06	2.64E-05	-1.31	0.192	6	-3.22E-05	5.89E-04	-0.97	0.335		
	6	2.12E-07	2.38E-05	0.16	0.875	7	2.00E-06	3.64E-04	0.10	0.922		
	7	-8.10E-07	2.16E-05	-0.66	0.508	8	-1.33E-05	1.87E-04	-1.26	0.209		
	8	1.01E-06	1.98E-05	0.91	0.366	9	6.68E-06	3.18E-04	0.37	0.711		
	9	-1.23E-06	2.52E-05	-0.87	0.387	10	-1.50E-05	2.06E-04	-1.29	0.197		
Depth	0	-1.15E-06	9.81E-07	-20.69	0.000	***	1	-7.83E-06	1.48E-05	-9.33	0.000	***
	1	-2.18E-07	3.92E-07	-9.86	0.000	***	2	-1.71E-06	9.17E-06	-3.29	0.001	***
	2	-8.30E-08	2.50E-07	-5.86	0.000	***	3	-7.33E-07	4.22E-06	-3.07	0.002	***
	3	-7.27E-08	3.54E-07	-3.64	0.000	***	4	-1.36E-07	5.22E-06	-0.46	0.646	
	4	-9.40E-10	2.24E-07	-0.07	0.941	5	2.87E-08	4.68E-06	0.11	0.914		
	5	-2.43E-08	2.87E-07	-1.50	0.136	6	2.22E-08	2.60E-06	0.15	0.880		
	6	1.74E-08	5.24E-07	0.59	0.558	7	-5.08E-08	6.76E-06	-0.13	0.894		
	7	-6.41E-08	4.77E-07	-2.38	0.018	**	8	1.13E-06	9.12E-06	2.20	0.029	**
	8	-5.85E-09	3.92E-07	-0.26	0.792	9	4.92E-08	8.03E-06	0.11	0.914		
	9	-2.09E-09	4.02E-07	-0.09	0.927	10	-2.73E-08	7.01E-06	1.10	0.272		



## A. SUPPLEMENT FOR CHAPTER 4

**Table A.13:** Adaptive Lasso Coefficients LG Chemicals (051910 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.118	0.071	-29.46	0.000	***	1	8.934	3.918	40.34	0.000	***
	2	-0.065	0.047	-24.63	0.000	***	2	-9.710	4.681	-36.70	0.000	***
	3	-0.043	0.033	-23.01	0.000	***	3	-4.307	2.217	-34.37	0.000	***
	4	-0.027	0.028	-17.32	0.000	***	4	-3.600	1.754	-36.31	0.000	***
	5	-0.017	0.024	-12.69	0.000	***	5	-2.293	1.448	-28.01	0.000	***
	6	-0.013	0.026	-8.42	0.000	***	6	-1.544	1.167	-23.41	0.000	***
	7	-0.007	0.016	-7.95	0.000	***	7	-0.993	0.943	-18.62	0.000	***
	8	-0.005	0.013	-7.11	0.000	***	8	-0.648	0.714	-16.07	0.000	***
	9	-0.004	0.015	-5.11	0.000	***	9	-0.308	0.492	-11.08	0.000	***
	10	-0.002	0.010	-3.63	0.000	***	10	-0.124	0.342	-6.42	0.000	***
Trades	0	1.50E-03	1.09E-03	24.36	0.000	***	1	0.242	0.106	40.57	0.000	***
	1	3.36E-04	2.68E-04	22.19	0.000	***	2	0.063	0.046	24.59	0.000	***
	2	2.92E-04	2.08E-04	24.78	0.000	***	3	0.069	0.018	66.33	0.000	***
	3	1.52E-04	1.55E-04	17.37	0.000	***	4	0.050	0.018	49.47	0.000	***
	4	1.01E-04	1.37E-04	12.99	0.000	***	5	0.041	0.016	44.44	0.000	***
	5	6.77E-05	1.16E-04	10.32	0.000	***	6	0.034	0.016	36.99	0.000	***
	6	4.32E-05	7.77E-05	9.84	0.000	***	7	0.030	0.014	37.21	0.000	***
	7	3.63E-05	7.61E-05	8.44	0.000	***	8	0.027	0.014	33.23	0.000	***
	8	2.80E-05	9.39E-05	5.28	0.000	***	9	0.025	0.014	31.51	0.000	***
	9	1.83E-05	4.63E-05	7.00	0.000	***	10	0.028	0.013	37.15	0.000	***
Durations	0	4.98E-06	4.97E-06	17.74	0.000	***	1	4.06E-06	9.51E-05	0.76	0.451	
	1	2.83E-07	3.97E-06	1.26	0.208	2	4.86E-06	7.38E-05	1.16	0.245		
	2	-3.92E-08	4.29E-06	-0.16	0.872	3	5.98E-07	4.78E-05	0.22	0.825		
	3	-1.81E-07	2.44E-06	-1.31	0.190	4	2.44E-06	4.84E-05	0.89	0.373		
	4	-1.44E-07	2.78E-06	-0.91	0.361	5	1.48E-06	4.51E-05	0.58	0.561		
	5	-3.15E-07	2.87E-06	-1.94	0.053	6	1.32E-06	2.75E-05	0.85	0.398		
	6	-1.25E-07	2.79E-06	-0.79	0.428	7	2.02E-07	3.25E-05	0.11	0.913		
	7	-1.25E-07	1.93E-06	-1.14	0.254	8	-1.38E-06	2.84E-05	-0.86	0.391		
	8	-2.99E-07	1.93E-06	-2.75	0.006	9	1.02E-07	2.50E-05	0.07	0.942		
	9	-2.90E-07	4.00E-06	-1.28	0.200	10	-1.48E-06	2.26E-05	-1.16	0.249		
10	8.85E-08	2.05E-06	0.76	0.447								
Spreads	0	2.00E-05	4.54E-05	7.80	0.000	***	1	2.64E-04	8.97E-04	5.21	0.000	***
	1	6.15E-06	2.80E-05	3.88	0.000	***	2	7.20E-05	3.40E-04	3.75	0.000	***
	2	1.67E-06	2.17E-05	1.36	0.175	3	6.05E-05	2.98E-04	3.59	0.000	***	
	3	3.52E-06	2.06E-05	3.03	0.003	***	4	1.30E-05	1.66E-04	1.39	0.166	
	4	2.11E-06	1.91E-05	1.96	0.051	5	2.92E-05	2.24E-04	2.31	0.021	**	
	5	1.05E-06	1.32E-05	1.41	0.160	6	2.53E-07	2.31E-04	0.02	0.985		
	6	5.28E-07	1.06E-05	0.89	0.376	7	1.93E-05	1.60E-04	2.14	0.033	**	
	7	3.21E-07	1.54E-05	0.37	0.712	8	7.96E-06	9.65E-05	1.46	0.146		
	8	1.07E-06	1.53E-05	1.23	0.219	9	9.60E-06	1.43E-04	1.19	0.237		
	9	-1.88E-07	1.58E-05	-0.21	0.833	10	1.35E-05	1.20E-04	1.99	0.048	**	
10	-4.85E-07	1.06E-05	-0.81	0.417								
Depth	0	-1.50E-06	1.31E-06	-20.19	0.000	***	1	-1.15E-05	1.62E-05	-12.55	0.000	***
	1	-3.04E-07	3.55E-07	-15.13	0.000	***	2	-1.84E-06	7.20E-06	-4.51	0.000	***
	2	-1.51E-07	2.64E-07	-10.14	0.000	***	3	-1.05E-06	6.02E-06	-3.07	0.002	***
	3	-5.85E-08	2.44E-07	-4.24	0.000	***	4	8.99E-07	7.71E-06	2.06	0.040	**
	4	-2.92E-08	1.53E-07	-3.37	0.001	***	5	8.33E-07	5.60E-06	2.63	0.009	***
	5	-1.14E-08	1.43E-07	-1.41	0.159	6	1.91E-06	7.92E-06	4.26	0.000	***	
	6	-1.87E-08	1.44E-07	-2.29	0.023	7	2.14E-06	8.05E-06	4.71	0.000	***	
	7	1.15E-08	2.16E-07	0.94	0.346	8	1.82E-06	6.41E-06	5.02	0.000	***	
	8	-1.48E-08	2.39E-07	-1.10	0.273	9	2.11E-06	6.39E-06	5.85	0.000	***	
	9	-6.28E-09	1.36E-07	-0.82	0.414	10	2.37E-06	6.61E-06	1.10	0.272		
10	1.69E-08	1.07E-07	2.80	0.005	***							

**Table A.14:** Adaptive Lasso Coefficients LG Electronics (066570 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value		lag	mean	stdev	T-stat	p-value	
Quote Revisions	1	-0.094	0.065	-25.51	0.000	***	1	8.667	5.490	27.93	0.000	***
	2	-0.062	0.050	-22.18	0.000	***	2	-12.474	5.919	-37.29	0.000	***
	3	-0.040	0.032	-21.85	0.000	***	3	-6.844	3.621	-33.44	0.000	***
	4	-0.028	0.034	-14.60	0.000	***	4	-5.666	2.704	-37.07	0.000	***
	5	-0.019	0.023	-14.61	0.000	***	5	-4.051	2.242	-31.96	0.000	***
	6	-0.014	0.023	-11.04	0.000	***	6	-2.955	1.845	-28.33	0.000	***
	7	-0.009	0.016	-10.13	0.000	***	7	-2.054	1.485	-24.47	0.000	***
	8	-0.007	0.017	-7.35	0.000	***	8	-1.311	1.179	-19.68	0.000	***
	9	-0.005	0.015	-6.48	0.000	***	9	-0.830	0.939	-15.64	0.000	***
	10	-0.003	0.011	-5.21	0.000	***	10	-0.409	0.638	-11.33	0.000	***
Trades	0	8.92E-04	7.17E-04	22.03	0.000	***	1	0.236	0.084	49.56	0.000	***
	1	2.02E-04	1.88E-04	19.05	0.000	***	2	0.073	0.043	30.06	0.000	***
	2	1.78E-04	1.44E-04	21.98	0.000	***	3	0.076	0.014	97.57	0.000	***
	3	1.04E-04	1.20E-04	15.31	0.000	***	4	0.058	0.014	74.62	0.000	***
	4	6.55E-05	8.80E-05	13.17	0.000	***	5	0.049	0.013	64.71	0.000	***
	5	4.83E-05	8.92E-05	9.58	0.000	***	6	0.042	0.013	57.66	0.000	***
	6	3.50E-05	6.59E-05	9.39	0.000	***	7	0.037	0.014	46.21	0.000	***
	7	2.78E-05	6.02E-05	8.17	0.000	***	8	0.035	0.013	47.11	0.000	***
	8	2.23E-05	5.62E-05	7.01	0.000	***	9	0.032	0.014	42.07	0.000	***
	9	1.69E-05	4.58E-05	6.52	0.000	***	10	0.036	0.013	47.94	0.000	***
Durations	0	3.53E-06	3.67E-06	17.01	0.000	***	1	-5.12E-07	7.36E-05	-0.12	0.902	
	1	3.27E-08	4.28E-06	0.14	0.893		2	1.09E-05	7.11E-05	2.70	0.007	***
	2	2.76E-07	2.58E-06	1.89	0.059	*	3	-1.34E-06	6.75E-05	-0.35	0.725	
	3	3.00E-08	2.17E-06	0.25	0.806		4	6.90E-06	4.42E-05	2.76	0.006	***
	4	-5.91E-08	5.73E-06	-0.18	0.855		5	3.88E-06	5.41E-05	1.27	0.205	
	5	3.62E-08	2.27E-06	0.28	0.778		6	1.79E-06	4.85E-05	0.65	0.515	
	6	3.20E-11	2.83E-06	0.00	1.000		7	3.80E-08	4.20E-05	0.02	0.987	
	7	6.85E-08	2.03E-06	0.60	0.551		8	-1.04E-07	3.56E-05	-0.05	0.959	
	8	-1.05E-07	1.44E-06	-1.29	0.200		9	1.07E-06	2.93E-05	0.65	0.518	
	9	-1.82E-07	1.69E-06	-1.90	0.059	*	10	-2.07E-07	2.05E-05	-0.18	0.859	
Spreads	0	2.34E-05	1.19E-04	3.48	0.001	***	1	8.68E-05	3.31E-03	0.46	0.644	
	1	1.32E-05	9.00E-05	2.59	0.010	***	2	4.71E-05	1.59E-03	0.52	0.602	
	2	4.36E-06	7.29E-05	1.06	0.292		3	1.37E-04	1.28E-03	1.88	0.061	*
	3	1.93E-05	1.13E-04	3.01	0.003	***	4	-3.45E-05	1.48E-03	-0.41	0.680	
	4	8.03E-07	5.00E-05	0.28	0.777		5	2.76E-04	1.88E-03	2.60	0.010	***
	5	1.98E-06	3.84E-05	0.91	0.364		6	8.88E-05	8.46E-04	1.85	0.065	*
	6	2.61E-06	3.63E-05	1.27	0.206		7	1.17E-04	8.10E-04	2.55	0.011	**
	7	5.69E-06	4.23E-05	2.38	0.018	**	8	5.73E-05	4.64E-04	2.18	0.030	**
	8	2.77E-06	3.13E-05	1.57	0.118		9	5.10E-05	4.02E-04	2.24	0.026	**
	9	-2.50E-07	3.79E-05	-0.12	0.907		10	3.80E-05	5.16E-04	1.30	0.194	
Depth	0	-3.80E-07	4.39E-07	-15.34	0.000	***	1	-6.45E-06	8.21E-06	-13.90	0.000	***
	1	-6.48E-08	1.08E-07	-10.63	0.000	***	2	-1.61E-06	4.41E-06	-6.47	0.000	***
	2	-3.09E-08	8.62E-08	-6.33	0.000	***	3	-7.99E-07	3.02E-06	-4.68	0.000	***
	3	-1.45E-08	5.22E-08	-4.91	0.000	***	4	-1.04E-07	3.13E-06	-0.59	0.557	
	4	-9.38E-09	3.96E-08	-4.19	0.000	***	5	-2.83E-07	2.43E-06	-2.06	0.040	**
	5	-2.32E-09	2.58E-08	-1.59	0.113		6	4.00E-07	3.26E-06	2.17	0.031	**
	6	-1.04E-09	9.35E-08	-0.20	0.844		7	5.41E-07	3.08E-06	3.11	0.002	***
	7	-4.46E-09	1.08E-07	-0.73	0.466		8	2.78E-07	1.96E-06	2.51	0.013	**
	8	2.02E-09	5.12E-08	0.70	0.486		9	1.01E-06	4.04E-06	4.43	0.000	***
	9	-1.05E-09	4.90E-08	-0.38	0.705		10	9.63E-07	3.42E-06	1.10	0.272	
10	-1.26E-10	3.03E-08	-0.07	0.942								

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.15:** Adaptive Lasso Coefficients Hana Financial Group (086790 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.143	0.071	-35.66	0.000	***	1	14.206	5.196	48.37	0.000	***
	2	-0.059	0.049	-21.20	0.000	***	2	-15.407	6.152	-44.31	0.000	***
	3	-0.033	0.039	-15.36	0.000	***	3	-4.400	2.926	-26.61	0.000	***
	4	-0.019	0.030	-11.12	0.000	***	4	-3.114	1.970	-27.96	0.000	***
	5	-0.010	0.025	-6.98	0.000	***	5	-1.340	1.468	-16.15	0.000	***
	6	-0.008	0.021	-6.48	0.000	***	6	-0.826	1.129	-12.95	0.000	***
	7	-0.006	0.021	-5.11	0.000	***	7	-0.427	0.779	-9.70	0.000	***
	8	-0.004	0.016	-3.83	0.000	***	8	-0.262	0.622	-7.46	0.000	***
	9	-0.002	0.012	-2.24	0.026	**	9	-0.116	0.337	-6.06	0.000	***
	10	-0.002	0.012	-2.71	0.007	***	10	-0.081	0.311	-4.58	0.000	***
Trades	0	1.57E-03	1.02E-03	27.21	0.000	***	1	0.311	0.088	62.27	0.000	***
	1	2.36E-04	2.51E-04	16.67	0.000	***	2	0.060	0.048	22.26	0.000	***
	2	1.75E-04	1.89E-04	16.37	0.000	***	3	0.065	0.018	64.26	0.000	***
	3	5.94E-05	1.07E-04	9.80	0.000	***	4	0.040	0.017	42.56	0.000	***
	4	3.36E-05	1.13E-04	5.26	0.000	***	5	0.033	0.015	38.83	0.000	***
	5	1.47E-05	5.84E-05	4.45	0.000	***	6	0.028	0.014	35.62	0.000	***
	6	1.37E-05	5.66E-05	4.29	0.000	***	7	0.024	0.014	29.80	0.000	***
	7	3.68E-06	2.68E-05	2.43	0.016	**	8	0.022	0.013	29.70	0.000	***
	8	-3.02E-07	2.61E-05	-0.20	0.838		9	0.020	0.014	25.00	0.000	***
	9	3.26E-07	2.14E-05	0.27	0.788		10	0.026	0.014	32.58	0.000	***
10	-3.33E-07	3.12E-05	-0.19	0.850								
Durations	0	6.62E-06	5.43E-06	21.58	0.000	***	1	6.54E-06	1.78E-04	0.65	0.516	
	1	-6.27E-08	7.00E-06	-0.16	0.874		2	7.70E-06	1.12E-04	1.21	0.226	
	2	-3.06E-07	4.25E-06	-1.27	0.204		3	5.67E-06	1.02E-04	0.99	0.325	
	3	3.21E-07	4.60E-06	1.23	0.219		4	-2.07E-06	6.02E-05	-0.61	0.544	
	4	-5.48E-08	4.01E-06	-0.24	0.809		5	4.00E-06	6.38E-05	1.11	0.268	
	5	3.30E-08	4.04E-06	0.14	0.885		6	-6.01E-06	7.21E-05	-1.48	0.141	
	6	-7.31E-08	3.16E-06	-0.41	0.683		7	4.25E-06	5.50E-05	1.37	0.173	
	7	-4.02E-08	2.19E-06	-0.32	0.746		8	-1.64E-06	4.51E-05	-0.64	0.520	
	8	-1.29E-07	1.85E-06	-1.23	0.219		9	-1.42E-06	3.02E-05	-0.83	0.407	
	9	-1.88E-07	3.55E-06	-0.94	0.350		10	6.42E-07	3.15E-05	0.36	0.718	
10	1.29E-07	2.53E-06	0.90	0.369								
Spreads	0	8.53E-05	1.64E-04	9.23	0.000	***	1	1.11E-03	4.35E-03	4.52	0.000	***
	1	2.96E-05	1.11E-04	4.72	0.000	***	2	1.92E-04	1.83E-03	1.85	0.065	*
	2	1.45E-05	9.62E-05	2.67	0.008	***	3	1.47E-04	1.56E-03	1.66	0.097	*
	3	6.71E-06	7.39E-05	1.60	0.110		4	2.02E-04	1.29E-03	2.78	0.006	***
	4	-4.48E-06	6.26E-05	-1.27	0.206		5	1.16E-04	1.05E-03	1.96	0.051	*
	5	2.22E-06	6.64E-05	0.59	0.555		6	1.91E-05	6.78E-04	0.50	0.618	
	6	4.53E-06	5.96E-05	1.34	0.180		7	-1.62E-05	9.38E-04	-0.31	0.761	
	7	4.32E-07	4.80E-05	0.16	0.874		8	4.76E-05	5.14E-04	1.64	0.102	
	8	-5.32E-06	5.44E-05	-1.73	0.085	*	9	3.06E-05	5.12E-04	1.05	0.292	
	9	-4.30E-06	4.32E-05	-1.76	0.079	*	10	2.28E-05	5.31E-04	0.76	0.447	
10	5.28E-06	4.19E-05	2.23	0.027	**							
Depth	0	-7.91E-07	6.74E-07	-20.78	0.000	***	1	-8.83E-06	1.25E-05	-12.47	0.000	***
	1	-1.05E-07	1.69E-07	-11.04	0.000	***	2	3.18E-07	3.93E-06	1.43	0.154	
	2	-2.70E-08	8.96E-08	-5.33	0.000	***	3	1.28E-07	3.33E-06	0.68	0.497	
	3	-1.83E-08	1.34E-07	-2.40	0.017	**	4	8.77E-07	4.05E-06	3.83	0.000	***
	4	-8.49E-09	1.24E-07	-1.21	0.228		5	6.26E-07	3.11E-06	3.56	0.000	***
	5	-5.88E-09	1.07E-07	-0.97	0.330		6	1.12E-06	3.77E-06	5.24	0.000	***
	6	-9.00E-09	1.65E-07	-0.96	0.336		7	1.05E-06	4.02E-06	4.64	0.000	***
	7	5.51E-09	1.64E-07	0.59	0.553		8	1.04E-06	3.90E-06	4.73	0.000	***
	8	2.30E-08	2.54E-07	1.60	0.110		9	7.18E-07	2.87E-06	4.43	0.000	***
	9	3.91E-09	1.45E-07	0.48	0.634		10	1.74E-06	5.01E-06	1.10	0.272	
10	1.30E-08	1.68E-07	1.37	0.171								

**Table A.16:** Adaptive Lasso Coefficients Hyundai Engineering & Construction (000720 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value		lag	mean	stdev	T-stat	p-value	
Quote Revisions	1	-0.132	0.069	-33.76	0.000	***	1	11.721	4.019	51.59	0.000	***
	2	-0.075	0.046	-28.80	0.000	***	2	-13.265	5.228	-44.89	0.000	***
	3	-0.052	0.036	-25.77	0.000	***	3	-6.180	3.079	-35.52	0.000	***
	4	-0.032	0.035	-16.40	0.000	***	4	-4.829	2.312	-36.96	0.000	***
	5	-0.020	0.029	-12.68	0.000	***	5	-3.084	1.913	-28.52	0.000	***
	6	-0.015	0.027	-9.68	0.000	***	6	-2.020	1.558	-22.93	0.000	***
	7	-0.012	0.021	-10.12	0.000	***	7	-1.237	1.257	-17.40	0.000	***
	8	-0.007	0.021	-5.74	0.000	***	8	-0.714	0.913	-13.83	0.000	***
	9	-0.005	0.014	-6.13	0.000	***	9	-0.366	0.637	-10.15	0.000	***
	10	-0.002	0.011	-3.60	0.000	***	10	-0.114	0.355	-5.70	0.000	***
Trades	0	1.31E-03	8.24E-04	28.06	0.000	***	1	0.260	0.085	54.01	0.000	***
	1	2.99E-04	2.17E-04	24.39	0.000	***	2	0.068	0.047	25.69	0.000	***
	2	2.60E-04	1.79E-04	25.65	0.000	***	3	0.074	0.016	81.05	0.000	***
	3	1.31E-04	1.49E-04	15.53	0.000	***	4	0.051	0.014	65.68	0.000	***
	4	8.48E-05	1.14E-04	13.12	0.000	***	5	0.042	0.012	60.97	0.000	***
	5	4.93E-05	8.78E-05	9.93	0.000	***	6	0.035	0.013	47.55	0.000	***
	6	4.22E-05	9.63E-05	7.76	0.000	***	7	0.030	0.013	41.50	0.000	***
	7	3.04E-05	8.28E-05	6.50	0.000	***	8	0.027	0.011	42.32	0.000	***
	8	2.47E-05	7.74E-05	5.65	0.000	***	9	0.025	0.012	35.20	0.000	***
	9	1.22E-05	3.88E-05	5.56	0.000	***	10	0.028	0.011	44.14	0.000	***
Durations	0	4.64E-06	3.96E-06	20.73	0.000	***	1	-3.67E-06	1.03E-04	-0.63	0.528	
	1	3.51E-07	5.09E-06	1.22	0.223		2	6.35E-06	8.94E-05	1.26	0.210	
	2	4.74E-07	6.31E-06	1.33	0.184		3	-1.56E-06	1.28E-04	-0.22	0.829	
	3	-1.56E-07	3.39E-06	-0.81	0.416		4	1.07E-05	9.22E-05	2.05	0.041	**
	4	2.42E-07	3.11E-06	1.38	0.170		5	2.32E-06	6.88E-05	0.60	0.552	
	5	-3.45E-08	3.43E-06	-0.18	0.859		6	5.84E-06	5.11E-05	2.02	0.044	**
	6	-2.75E-07	3.35E-06	-1.45	0.147		7	2.51E-06	5.21E-05	0.85	0.394	
	7	-3.20E-07	3.10E-06	-1.82	0.069	*	8	1.20E-06	3.74E-05	0.57	0.571	
	8	1.17E-07	1.65E-06	1.26	0.210		9	-4.05E-06	3.96E-05	-1.81	0.072	*
	9	-1.14E-07	2.26E-06	-0.89	0.372		10	1.76E-07	2.72E-05	0.11	0.909	
Spreads	0	3.13E-05	1.01E-04	5.50	0.000	***	1	4.36E-04	2.69E-03	2.86	0.004	***
	1	1.02E-05	8.90E-05	2.02	0.044	**	2	1.32E-04	1.03E-03	2.26	0.024	**
	2	1.52E-05	6.01E-05	4.48	0.000	***	3	7.19E-05	1.12E-03	1.13	0.258	
	3	3.29E-06	4.62E-05	1.26	0.209		4	1.19E-04	7.69E-04	2.74	0.006	***
	4	5.49E-06	4.94E-05	1.97	0.050	*	5	1.13E-04	7.34E-04	2.73	0.007	***
	5	8.20E-06	4.57E-05	3.18	0.002	***	6	5.83E-05	5.35E-04	1.93	0.055	*
	6	-5.08E-07	5.60E-05	-0.16	0.873		7	1.31E-04	9.30E-04	2.49	0.013	**
	7	5.25E-07	2.60E-05	0.36	0.722		8	5.24E-05	6.55E-04	1.42	0.158	
	8	1.05E-06	3.60E-05	0.52	0.606		9	3.40E-05	4.67E-04	1.29	0.198	
	9	-2.79E-06	3.73E-05	-1.32	0.187		10	6.69E-05	6.24E-04	1.90	0.059	*
Depth	0	-9.18E-07	8.15E-07	-19.94	0.000	***	1	-1.14E-05	1.33E-05	-15.21	0.000	***
	1	-1.61E-07	2.13E-07	-13.38	0.000	***	2	-7.88E-07	4.87E-06	-2.86	0.004	***
	2	-7.69E-08	1.53E-07	-8.87	0.000	***	3	7.49E-07	7.03E-06	1.88	0.060	*
	3	-4.02E-08	1.32E-07	-5.37	0.000	***	4	8.18E-07	4.86E-06	2.98	0.003	***
	4	-1.03E-08	7.80E-08	-2.35	0.020	**	5	1.71E-06	6.51E-06	4.65	0.000	***
	5	1.47E-10	8.54E-08	0.03	0.976		6	2.07E-06	5.88E-06	6.24	0.000	***
	6	3.22E-09	1.17E-07	0.49	0.628		7	2.41E-06	6.69E-06	6.36	0.000	***
	7	3.01E-09	1.24E-07	0.43	0.668		8	1.94E-06	5.75E-06	5.98	0.000	***
	8	3.48E-10	9.64E-08	0.06	0.949		9	3.07E-06	8.09E-06	6.72	0.000	***
	9	-2.87E-09	8.96E-08	-0.57	0.572		10	4.33E-06	7.72E-06	1.10	0.272	
10	2.70E-08	1.36E-07	3.52	0.000	***							

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.17:** Adaptive Lasso Coefficients Samsung C&T Corporation (000830 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation						Trade Equation					
	lag	mean	stdev	T-stat	p-value		lag	mean	stdev	T-stat	p-value	
Quote Revisions	1	-0.142	0.052	-48.01	0.000	***	1	7.552	1.865	71.63	0.000	***
	2	-0.068	0.028	-42.46	0.000	***	2	-5.949	2.261	-46.55	0.000	***
	3	-0.041	0.022	-33.27	0.000	***	3	-1.715	1.181	-25.69	0.000	***
	4	-0.023	0.019	-21.32	0.000	***	4	-1.213	0.988	-21.72	0.000	***
	5	-0.015	0.017	-15.62	0.000	***	5	-0.470	0.689	-12.07	0.000	***
	6	-0.010	0.015	-11.86	0.000	***	6	-0.317	0.546	-10.28	0.000	***
	7	-0.005	0.012	-7.13	0.000	***	7	-0.131	0.315	-7.33	0.000	***
	8	-0.004	0.009	-7.07	0.000	***	8	-0.092	0.277	-5.87	0.000	***
	9	-0.003	0.008	-5.63	0.000	***	9	-0.052	0.187	-4.89	0.000	***
	10	-0.002	0.010	-3.27	0.001	***	10	-0.024	0.135	-3.09	0.002	***
Trades	0	2.71E-03	1.72E-03	27.87	0.000	***	1	0.268	0.090	52.72	0.000	***
	1	5.09E-04	4.61E-04	19.54	0.000	***	2	0.063	0.038	29.62	0.000	***
	2	3.99E-04	3.84E-04	18.35	0.000	***	3	0.061	0.019	57.37	0.000	***
	3	1.36E-04	2.05E-04	11.72	0.000	***	4	0.038	0.015	44.35	0.000	***
	4	9.05E-05	1.74E-04	9.20	0.000	***	5	0.030	0.015	36.55	0.000	***
	5	4.51E-05	1.14E-04	7.00	0.000	***	6	0.025	0.014	31.67	0.000	***
	6	2.55E-05	1.06E-04	4.27	0.000	***	7	0.021	0.014	26.74	0.000	***
	7	1.32E-05	6.05E-05	3.86	0.000	***	8	0.020	0.014	25.76	0.000	***
	8	1.27E-05	6.63E-05	3.40	0.001	***	9	0.019	0.014	24.44	0.000	***
	9	-3.06E-06	5.14E-05	-1.05	0.293		10	0.022	0.014	28.64	0.000	***
Durations	0	7.25E-06	5.76E-06	22.28	0.000	***	1	2.66E-08	5.07E-05	0.01	0.993	
	1	1.05E-07	2.81E-06	0.66	0.511		2	1.51E-06	4.17E-05	0.64	0.523	
	2	-1.13E-08	1.68E-06	-0.12	0.905		3	9.31E-07	2.19E-05	0.75	0.453	
	3	1.02E-07	1.62E-06	1.11	0.266		4	-1.04E-07	7.87E-06	-0.23	0.816	
	4	-1.45E-08	1.38E-06	-0.19	0.852		5	9.82E-07	1.46E-05	1.19	0.234	
	5	-6.16E-08	9.79E-07	-1.11	0.266		6	3.12E-07	4.92E-06	1.12	0.263	
	6	-8.24E-09	9.67E-07	-0.15	0.880		7	1.21E-07	1.99E-06	1.07	0.284	
	7	-9.50E-08	9.71E-07	-1.73	0.085	*	8	-2.21E-07	4.02E-06	-0.97	0.332	
	8	-3.87E-08	8.69E-07	-0.79	0.431		9	7.53E-08	1.58E-06	0.84	0.401	
	9	3.48E-08	8.11E-07	0.76	0.448		10	-6.85E-07	5.77E-06	-2.10	0.037	**
Spreads	0	8.01E-06	2.13E-05	6.64	0.000	***	1	6.77E-05	2.68E-04	4.47	0.000	***
	1	3.34E-06	1.43E-05	4.13	0.000	***	2	1.81E-06	1.34E-04	0.24	0.811	
	2	1.59E-06	1.60E-05	1.76	0.079	*	3	1.95E-05	1.18E-04	2.93	0.004	***
	3	2.35E-06	9.85E-06	4.22	0.000	***	4	-1.78E-05	1.62E-04	-1.95	0.052	*
	4	2.63E-07	8.53E-06	0.55	0.585		5	3.65E-06	6.54E-05	0.99	0.325	
	5	1.84E-07	8.03E-06	0.41	0.685		6	2.78E-07	5.14E-05	0.10	0.924	
	6	5.42E-07	7.10E-06	1.35	0.178		7	-2.34E-06	4.82E-05	-0.86	0.391	
	7	2.04E-08	7.45E-06	0.05	0.961		8	-2.03E-07	7.04E-05	-0.05	0.959	
	8	-2.74E-07	6.41E-06	-0.75	0.451		9	2.15E-06	4.69E-05	0.81	0.419	
	9	3.19E-07	6.23E-06	0.90	0.366		10	8.96E-07	8.11E-05	0.20	0.845	
Depth	0	-4.12E-06	3.87E-06	-18.86	0.000	***	1	-8.66E-06	2.07E-05	-7.39	0.000	***
	1	-8.05E-07	1.61E-06	-8.87	0.000	***	2	-1.17E-06	9.71E-06	-2.13	0.034	**
	2	-4.38E-07	1.18E-06	-6.58	0.000	***	3	2.95E-07	1.14E-05	0.46	0.648	
	3	-1.10E-07	5.32E-07	-3.65	0.000	***	4	-1.83E-07	4.44E-06	-0.73	0.466	
	4	-5.92E-08	5.07E-07	-2.06	0.040	**	5	9.03E-07	9.15E-06	1.75	0.082	*
	5	-3.23E-08	4.00E-07	-1.43	0.154		6	1.69E-06	9.36E-06	3.19	0.002	***
	6	-3.01E-09	4.39E-07	-0.12	0.904		7	1.01E-06	5.81E-06	3.08	0.002	***
	7	3.22E-09	7.80E-07	0.07	0.942		8	1.13E-06	6.45E-06	3.10	0.002	***
	8	-3.88E-08	6.35E-07	-1.08	0.281		9	3.15E-07	4.75E-06	1.17	0.241	
	9	-8.86E-09	5.74E-07	-0.27	0.785		10	1.02E-06	6.64E-06	1.10	0.272	
10	3.22E-08	4.52E-07	1.26	0.209								

**Table A.18:** Adaptive Lasso Coefficients LG Corporation (003550 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.131	0.052	-44.85	0.000	***	1	13.214	4.109	56.89	0.000	***
	2	-0.059	0.035	-29.51	0.000	***	2	-13.297	5.246	-44.84	0.000	***
	3	-0.033	0.025	-23.89	0.000	***	3	-4.545	2.591	-31.04	0.000	***
	4	-0.019	0.025	-13.42	0.000	***	4	-3.309	2.131	-27.47	0.000	***
	5	-0.011	0.018	-10.52	0.000	***	5	-1.535	1.475	-18.41	0.000	***
	6	-0.006	0.016	-6.47	0.000	***	6	-0.868	1.071	-14.34	0.000	***
	7	-0.003	0.010	-5.01	0.000	***	7	-0.434	0.767	-10.02	0.000	***
	8	-0.003	0.012	-3.74	0.000	***	8	-0.240	0.588	-7.21	0.000	***
	9	-0.002	0.012	-2.61	0.010	***	9	-0.099	0.333	-5.24	0.000	***
	10	-0.002	0.011	-3.03	0.003	***	10	-0.030	0.251	-2.08	0.038	**
Trades	0	1.72E-03	9.20E-04	33.10	0.000	***	1	0.283	0.099	50.70	0.000	***
	1	3.74E-04	3.04E-04	21.74	0.000	***	2	0.064	0.048	23.41	0.000	***
	2	2.77E-04	2.02E-04	24.20	0.000	***	3	0.068	0.016	73.67	0.000	***
	3	9.42E-05	1.22E-04	13.62	0.000	***	4	0.043	0.016	47.16	0.000	***
	4	5.05E-05	9.49E-05	9.42	0.000	***	5	0.035	0.014	44.72	0.000	***
	5	2.40E-05	7.36E-05	5.77	0.000	***	6	0.027	0.014	32.77	0.000	***
	6	1.63E-05	6.26E-05	4.62	0.000	***	7	0.024	0.013	33.24	0.000	***
	7	9.83E-06	5.29E-05	3.29	0.001	***	8	0.022	0.012	31.49	0.000	***
	8	3.90E-06	3.75E-05	1.84	0.067	*	9	0.019	0.013	26.46	0.000	***
	9	6.19E-06	3.85E-05	2.84	0.005	***	10	0.025	0.012	35.59	0.000	***
10	-1.13E-06	2.49E-05	-0.80	0.423								
Durations	0	6.21E-06	4.78E-06	22.97	0.000	***	1	5.29E-06	2.42E-04	0.39	0.699	
	1	1.04E-07	5.34E-06	0.34	0.731		2	-4.22E-06	1.27E-04	-0.59	0.558	
	2	-3.33E-08	2.88E-06	-0.20	0.838		3	3.10E-06	7.99E-05	0.69	0.493	
	3	-1.66E-08	3.18E-06	-0.09	0.927		4	5.88E-07	4.66E-05	0.22	0.823	
	4	2.99E-07	2.60E-06	2.03	0.043	**	5	-2.18E-06	4.08E-05	-0.94	0.346	
	5	3.47E-08	2.73E-06	0.22	0.822		6	2.90E-06	4.67E-05	1.10	0.272	
	6	-1.29E-07	2.01E-06	-1.14	0.255		7	1.91E-07	3.29E-05	0.10	0.918	
	7	9.76E-08	2.04E-06	0.85	0.399		8	-9.24E-07	3.42E-05	-0.48	0.634	
	8	4.90E-08	1.46E-06	0.60	0.552		9	-1.01E-06	2.99E-05	-0.60	0.550	
	9	-3.84E-08	2.15E-06	-0.32	0.752		10	1.98E-06	3.10E-05	1.13	0.260	
10	7.14E-08	1.43E-06	0.88	0.378								
Spreads	0	4.12E-05	9.20E-05	7.92	0.000	***	1	7.01E-04	2.14E-03	5.78	0.000	***
	1	1.19E-05	5.67E-05	3.72	0.000	***	2	2.43E-04	1.28E-03	3.36	0.001	***
	2	8.72E-06	4.34E-05	3.55	0.000	***	3	6.33E-05	9.73E-04	1.15	0.251	
	3	2.10E-06	3.19E-05	1.17	0.244		4	9.92E-05	5.36E-04	3.27	0.001	***
	4	2.25E-06	4.50E-05	0.88	0.377		5	-1.56E-05	4.51E-04	-0.61	0.541	
	5	1.19E-06	3.09E-05	0.68	0.495		6	5.99E-05	5.55E-04	1.91	0.057	*
	6	1.14E-06	3.23E-05	0.63	0.532		7	2.14E-06	3.29E-04	0.12	0.908	
	7	8.28E-07	2.87E-05	0.51	0.610		8	-9.61E-06	3.17E-04	-0.54	0.592	
	8	-2.21E-06	3.69E-05	-1.06	0.290		9	-1.12E-06	2.62E-04	-0.08	0.940	
	9	3.61E-07	1.88E-05	0.34	0.734		10	9.72E-06	8.54E-05	2.01	0.045	**
10	-2.44E-06	1.79E-05	-2.41	0.016	**							
Depth	0	-1.40E-06	9.69E-07	-25.52	0.000	***	1	-9.81E-06	1.49E-05	-11.62	0.000	***
	1	-2.90E-07	3.88E-07	-13.25	0.000	***	2	1.86E-07	6.27E-06	0.53	0.600	
	2	-1.01E-07	2.42E-07	-7.41	0.000	***	3	4.02E-07	4.16E-06	1.71	0.089	*
	3	-1.06E-08	5.73E-07	-0.33	0.744		4	7.41E-07	8.73E-06	1.50	0.134	
	4	-4.14E-08	3.20E-07	-2.29	0.023	**	5	1.96E-06	9.26E-06	3.75	0.000	***
	5	7.41E-09	2.88E-07	0.45	0.650		6	1.79E-06	8.50E-06	3.73	0.000	***
	6	-9.79E-09	4.43E-07	-0.39	0.696		7	2.16E-06	7.92E-06	4.83	0.000	***
	7	-1.61E-08	4.36E-07	-0.65	0.515		8	1.63E-06	7.21E-06	4.00	0.000	***
	8	2.48E-08	3.29E-07	1.33	0.184		9	1.67E-06	9.55E-06	3.09	0.002	***
	9	1.27E-08	1.83E-07	1.23	0.219		10	1.43E-06	5.60E-06	1.10	0.272	
10	1.43E-08	2.10E-07	1.21	0.227								

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.19:** Adaptive Lasso Coefficients Samsung Electro-Mechanics (009150 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.113	0.076	-26.47	0.000	***	1	9.444	5.380	31.06	0.000	***
	2	-0.066	0.048	-24.48	0.000	***	2	-14.598	8.028	-32.17	0.000	***
	3	-0.047	0.048	-17.54	0.000	***	3	-6.871	3.549	-34.25	0.000	***
	4	-0.032	0.031	-18.32	0.000	***	4	-5.784	2.914	-35.12	0.000	***
	5	-0.018	0.025	-12.38	0.000	***	5	-3.640	2.204	-29.22	0.000	***
	6	-0.013	0.025	-9.18	0.000	***	6	-2.540	1.758	-25.56	0.000	***
	7	-0.010	0.021	-7.99	0.000	***	7	-1.521	1.328	-20.26	0.000	***
	8	-0.006	0.016	-6.87	0.000	***	8	-0.887	1.042	-15.06	0.000	***
	9	-0.005	0.017	-4.94	0.000	***	9	-0.468	0.704	-11.77	0.000	***
	10	-0.003	0.011	-5.24	0.000	***	10	-0.184	0.451	-7.21	0.000	***
Trades	0	1.01E-03	7.62E-04	23.55	0.000	***	1	0.257	0.108	42.04	0.000	***
	1	2.13E-04	2.29E-04	16.48	0.000	***	2	0.068	0.052	23.38	0.000	***
	2	2.24E-04	2.06E-04	19.24	0.000	***	3	0.081	0.019	76.52	0.000	***
	3	1.29E-04	1.31E-04	17.42	0.000	***	4	0.057	0.017	58.93	0.000	***
	4	7.24E-05	9.72E-05	13.17	0.000	***	5	0.048	0.016	54.47	0.000	***
	5	5.14E-05	9.76E-05	9.32	0.000	***	6	0.039	0.016	43.89	0.000	***
	6	3.73E-05	7.42E-05	8.90	0.000	***	7	0.034	0.015	39.70	0.000	***
	7	2.08E-05	5.60E-05	6.58	0.000	***	8	0.031	0.016	33.19	0.000	***
	8	2.10E-05	5.83E-05	6.38	0.000	***	9	0.029	0.016	31.69	0.000	***
	9	1.64E-05	4.42E-05	6.59	0.000	***	10	0.031	0.016	35.18	0.000	***
Durations	0	3.66E-06	3.49E-06	18.55	0.000	***	1	1.01E-06	1.31E-04	0.14	0.892	
	1	-1.64E-07	5.51E-06	-0.53	0.600		2	7.14E-06	1.10E-04	1.15	0.253	
	2	-3.32E-07	7.97E-06	-0.74	0.461		3	3.48E-06	9.55E-05	0.64	0.520	
	3	2.07E-07	5.10E-06	0.72	0.473		4	1.10E-06	9.09E-05	0.21	0.831	
	4	3.02E-07	3.70E-06	1.45	0.149		5	1.75E-06	7.70E-05	0.40	0.688	
	5	3.31E-07	2.77E-06	2.11	0.036	**	6	2.92E-07	6.21E-05	0.08	0.934	
	6	-1.05E-07	3.19E-06	-0.58	0.562		7	2.69E-06	4.79E-05	0.99	0.322	
	7	1.28E-07	2.33E-06	0.97	0.333		8	-1.21E-06	5.23E-05	-0.41	0.683	
	8	-8.56E-08	2.77E-06	-0.55	0.586		9	3.27E-06	3.37E-05	1.72	0.087	*
	9	-1.26E-07	2.29E-06	-0.97	0.331		10	-1.04E-06	3.36E-05	-0.55	0.585	
Spreads	0	3.09E-05	1.32E-04	4.16	0.000	***	1	1.31E-04	3.08E-03	0.75	0.453	
	1	4.39E-06	9.28E-05	0.84	0.403		2	3.55E-04	2.06E-03	3.05	0.002	***
	2	6.31E-06	1.01E-04	1.10	0.272		3	9.71E-05	2.34E-03	0.73	0.463	
	3	1.07E-05	7.99E-05	2.37	0.018	**	4	1.21E-04	1.60E-03	1.34	0.183	
	4	8.64E-06	8.51E-05	1.80	0.073	*	5	1.47E-04	1.47E-03	1.77	0.078	*
	5	-7.99E-08	6.83E-05	-0.02	0.984		6	1.50E-04	1.07E-03	2.48	0.014	**
	6	8.79E-07	6.76E-05	0.23	0.818		7	1.07E-05	8.15E-04	0.23	0.816	
	7	6.16E-06	9.02E-05	1.21	0.228		8	-1.90E-05	1.40E-03	-0.24	0.810	
	8	-1.48E-06	5.42E-05	-0.48	0.629		9	1.48E-04	1.95E-03	1.34	0.180	
	9	5.11E-06	3.96E-05	2.29	0.023	**	10	3.98E-05	2.92E-04	2.42	0.016	**
Depth	0	-6.39E-07	7.59E-07	-14.89	0.000	***	1	-1.48E-05	1.77E-05	-14.81	0.000	***
	1	-8.04E-08	1.83E-07	-7.79	0.000	***	2	-9.74E-07	5.48E-06	-3.14	0.002	***
	2	-5.71E-08	1.48E-07	-6.82	0.000	***	3	-1.15E-08	4.30E-06	-0.05	0.962	
	3	-2.91E-08	1.47E-07	-3.50	0.001	***	4	1.34E-06	7.72E-06	3.08	0.002	***
	4	-1.21E-08	7.23E-08	-2.97	0.003	***	5	1.14E-06	5.48E-06	3.69	0.000	***
	5	-3.91E-09	6.14E-08	-1.13	0.260		6	2.84E-06	1.06E-05	4.72	0.000	***
	6	-1.26E-09	3.82E-08	-0.58	0.561		7	2.54E-06	7.83E-06	5.74	0.000	***
	7	1.30E-09	2.66E-08	0.86	0.389		8	2.55E-06	7.01E-06	6.44	0.000	***
	8	7.81E-10	3.87E-08	0.36	0.721		9	2.75E-06	7.13E-06	6.81	0.000	***
	9	-4.36E-10	4.99E-08	-0.15	0.877		10	2.89E-06	6.65E-06	1.10	0.272	
10	7.79E-09	5.49E-08	2.51	0.013	**							

**Table A.20:** Adaptive Lasso Coefficients Samsung Heavy Industries (010140 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.118	0.052	-39.70	0.000	***	1	12.727	4.629	48.64	0.000	***
	2	-0.062	0.043	-25.88	0.000	***	2	-19.394	6.269	-54.73	0.000	***
	3	-0.035	0.034	-18.27	0.000	***	3	-9.125	3.997	-40.39	0.000	***
	4	-0.023	0.028	-14.32	0.000	***	4	-6.840	3.310	-36.56	0.000	***
	5	-0.016	0.040	-6.83	0.000	***	5	-4.153	2.676	-27.46	0.000	***
	6	-0.006	0.018	-5.84	0.000	***	6	-2.456	2.158	-20.13	0.000	***
	7	-0.006	0.022	-4.64	0.000	***	7	-1.449	1.638	-15.65	0.000	***
	8	-0.003	0.013	-3.56	0.000	***	8	-0.721	1.116	-11.43	0.000	***
	9	-0.001	0.014	-1.09	0.276	***	9	-0.360	0.804	-7.92	0.000	***
	10	-0.002	0.012	-2.96	0.003	***	10	-0.141	0.480	-5.21	0.000	***
Trades	0	9.82E-04	4.53E-04	38.32	0.000	***	1	0.285	0.069	72.88	0.000	***
	1	2.34E-04	2.13E-04	19.41	0.000	***	2	0.078	0.058	23.78	0.000	***
	2	1.76E-04	1.29E-04	24.12	0.000	***	3	0.081	0.017	84.64	0.000	***
	3	8.55E-05	1.06E-04	14.34	0.000	***	4	0.058	0.014	76.19	0.000	***
	4	6.52E-05	1.10E-04	10.52	0.000	***	5	0.045	0.013	60.94	0.000	***
	5	2.96E-05	6.70E-05	7.83	0.000	***	6	0.036	0.012	51.33	0.000	***
	6	2.74E-05	6.72E-05	7.22	0.000	***	7	0.030	0.012	45.46	0.000	***
	7	1.74E-05	4.73E-05	6.51	0.000	***	8	0.025	0.012	38.31	0.000	***
	8	7.12E-06	3.31E-05	3.80	0.000	***	9	0.024	0.011	39.52	0.000	***
	9	7.83E-06	3.47E-05	3.99	0.000	***	10	0.027	0.011	45.42	0.000	***
Durations	0	3.96E-06	3.04E-06	23.06	0.000	***	1	5.88E-06	7.61E-05	1.37	0.173	
	1	2.76E-07	6.64E-06	0.73	0.463	*	2	6.95E-06	1.26E-04	0.97	0.331	
	2	4.32E-07	4.53E-06	1.69	0.092	*	3	6.55E-06	1.34E-04	0.87	0.387	
	3	1.96E-07	4.68E-06	0.74	0.459	*	4	1.52E-05	1.10E-04	2.45	0.015	**
	4	2.18E-07	3.56E-06	1.08	0.280	*	5	6.70E-06	9.44E-05	1.26	0.210	
	5	2.80E-07	3.19E-06	1.55	0.121	*	6	1.42E-05	7.12E-05	3.52	0.000	***
	6	2.64E-08	3.47E-06	0.13	0.893	*	7	9.37E-06	6.49E-05	2.55	0.011	**
	7	-3.55E-08	3.16E-06	-0.20	0.842	*	8	1.07E-05	7.30E-05	2.60	0.010	***
	8	-1.40E-07	3.61E-06	-0.69	0.493	*	9	4.64E-06	6.44E-05	1.27	0.203	
	9	-1.15E-07	1.58E-06	-1.28	0.201	*	10	-2.69E-06	4.11E-05	-1.16	0.247	
Spreads	0	7.45E-05	3.12E-04	4.23	0.000	***	1	1.94E-04	1.02E-02	0.34	0.736	
	1	5.68E-05	2.19E-04	4.60	0.000	***	2	1.57E-04	5.40E-03	0.52	0.607	
	2	3.92E-05	1.99E-04	3.49	0.001	***	3	-6.41E-05	2.88E-03	-0.39	0.694	
	3	1.69E-05	1.94E-04	1.54	0.124	**	4	5.19E-04	3.52E-03	2.61	0.010	***
	4	2.99E-05	2.06E-04	2.56	0.011	**	5	1.51E-04	2.98E-03	0.89	0.372	
	5	-1.23E-05	1.83E-04	-1.18	0.237	*	6	5.45E-04	3.70E-03	2.61	0.010	***
	6	2.14E-06	1.11E-04	0.34	0.733	*	7	2.46E-05	2.20E-03	0.20	0.844	
	7	-1.68E-06	8.99E-05	-0.33	0.741	*	8	-4.35E-05	2.52E-03	-0.30	0.761	
	8	4.55E-06	9.23E-05	0.87	0.384	*	9	1.74E-04	1.37E-03	2.25	0.025	**
	9	7.85E-06	7.70E-05	1.80	0.072	*	10	-1.68E-06	1.17E-03	-0.03	0.980	
Depth	0	-3.54E-07	2.51E-07	-24.95	0.000	***	1	-1.70E-05	1.35E-05	-22.39	0.000	***
	1	-7.73E-08	1.18E-07	-11.56	0.000	***	2	-1.97E-06	5.67E-06	-6.13	0.000	***
	2	-1.85E-08	5.25E-08	-6.26	0.000	***	3	-2.78E-07	1.38E-06	-3.56	0.000	***
	3	-3.58E-09	3.89E-08	-1.63	0.104	**	4	2.48E-07	1.80E-06	2.43	0.016	**
	4	-3.19E-09	3.49E-08	-1.62	0.107	**	5	1.04E-06	3.54E-06	5.17	0.000	***
	5	-8.37E-10	2.66E-08	-0.56	0.578	**	6	1.45E-06	4.38E-06	5.85	0.000	***
	6	-7.91E-10	2.25E-08	-0.62	0.534	**	7	1.91E-06	4.47E-06	7.58	0.000	***
	7	5.82E-10	1.78E-08	0.58	0.563	**	8	2.10E-06	4.77E-06	7.79	0.000	***
	8	2.73E-09	3.35E-08	1.44	0.151	*	9	2.52E-06	4.75E-06	9.37	0.000	***
	9	5.08E-09	4.85E-08	1.85	0.065	*	10	2.68E-06	4.82E-06	1.10	0.272	
10	6.22E-09	2.41E-08	4.57	0.000	***							



## A. SUPPLEMENT FOR CHAPTER 4

**Table A.21:** Adaptive Lasso Coefficients S-Oil (010950 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.132	0.067	-35.03	0.000	***	1	10.422	4.839	38.10	0.000	***
	2	-0.056	0.032	-30.46	0.000	***	2	-9.499	4.804	-34.98	0.000	***
	3	-0.033	0.029	-20.40	0.000	***	3	-2.447	1.821	-23.78	0.000	***
	4	-0.016	0.018	-16.39	0.000	***	4	-1.796	1.422	-22.35	0.000	***
	5	-0.012	0.024	-9.25	0.000	***	5	-0.762	0.986	-13.67	0.000	***
	6	-0.007	0.013	-9.44	0.000	***	6	-0.464	0.781	-10.51	0.000	***
	7	-0.004	0.010	-7.38	0.000	***	7	-0.266	0.514	-9.16	0.000	***
	8	-0.003	0.010	-6.14	0.000	***	8	-0.179	0.420	-7.56	0.000	***
	9	-0.002	0.009	-4.76	0.000	***	9	-0.093	0.264	-6.26	0.000	***
	10	-0.001	0.007	-2.78	0.006	***	10	-0.048	0.194	-4.41	0.000	***
Trades	0	1.92E-03	1.27E-03	26.83	0.000	***	1	0.280	0.104	47.35	0.000	***
	1	3.56E-04	3.22E-04	19.54	0.000	***	2	0.065	0.047	24.64	0.000	***
	2	2.43E-04	2.47E-04	17.46	0.000	***	3	0.065	0.020	58.48	0.000	***
	3	7.48E-05	1.31E-04	10.06	0.000	***	4	0.039	0.020	34.36	0.000	***
	4	6.66E-05	1.42E-04	8.28	0.000	***	5	0.032	0.018	31.09	0.000	***
	5	3.12E-05	8.32E-05	6.64	0.000	***	6	0.025	0.017	25.51	0.000	***
	6	1.15E-05	4.13E-05	4.94	0.000	***	7	0.021	0.016	22.87	0.000	***
	7	1.10E-05	5.30E-05	3.66	0.000	***	8	0.020	0.016	21.72	0.000	***
	8	5.09E-06	2.45E-05	3.68	0.000	***	9	0.020	0.016	21.61	0.000	***
	9	6.67E-06	3.40E-05	3.47	0.001	***	10	0.021	0.016	22.77	0.000	***
10	-1.41E-06	3.42E-05	-0.73	0.467								
Durations	0	5.13E-06	4.09E-06	22.16	0.000	***	1	4.10E-06	1.25E-04	0.58	0.563	
	1	9.98E-08	4.11E-06	0.43	0.668		2	-4.05E-06	6.37E-05	-1.13	0.261	
	2	2.40E-07	2.59E-06	1.64	0.101		3	2.25E-07	3.41E-05	0.12	0.907	
	3	3.49E-07	4.24E-06	1.45	0.147		4	4.91E-07	4.48E-05	0.19	0.847	
	4	3.17E-07	5.27E-06	1.06	0.289		5	3.58E-06	4.77E-05	1.33	0.186	
	5	-2.15E-07	2.77E-06	-1.37	0.171		6	1.91E-06	4.28E-05	0.79	0.430	
	6	8.02E-08	2.06E-06	0.69	0.492		7	-4.72E-07	2.43E-05	-0.34	0.731	
	7	-4.06E-07	3.58E-06	-2.00	0.046	**	8	7.14E-07	1.56E-05	0.81	0.420	
	8	4.14E-08	1.03E-06	0.71	0.476		9	-1.83E-06	1.36E-05	-2.38	0.018	**
	9	1.09E-07	1.50E-06	1.28	0.201		10	3.50E-08	1.24E-05	0.05	0.960	
10	-2.82E-08	9.11E-07	-0.55	0.584								
Spreads	0	3.67E-05	1.35E-04	4.79	0.000	***	1	2.11E-04	9.24E-04	4.04	0.000	***
	1	1.14E-05	4.25E-05	4.73	0.000	***	2	7.29E-05	6.87E-04	1.88	0.062	*
	2	-1.43E-05	3.03E-04	-0.83	0.404		3	2.10E-05	3.81E-04	0.97	0.331	
	3	4.68E-06	3.50E-05	2.36	0.019	**	4	-1.16E-05	3.75E-04	-0.55	0.585	
	4	-5.67E-07	3.10E-05	-0.32	0.746		5	2.43E-05	4.12E-04	1.04	0.298	
	5	1.51E-06	2.27E-05	1.18	0.240		6	-3.93E-06	2.76E-04	-0.25	0.801	
	6	-1.02E-06	2.25E-05	-0.81	0.421		7	-3.45E-06	1.62E-04	-0.38	0.707	
	7	9.61E-07	1.58E-05	1.08	0.282		8	-1.17E-06	8.88E-05	-0.23	0.817	
	8	-2.00E-06	2.18E-05	-1.63	0.105		9	1.10E-05	2.18E-04	0.89	0.375	
	9	-2.09E-06	1.74E-05	-2.12	0.035	**	10	-6.20E-06	9.14E-05	-1.20	0.231	
10	-5.42E-07	2.25E-05	-0.43	0.671								
Depth	0	-1.73E-06	1.73E-06	-17.71	0.000	***	1	-4.99E-06	1.21E-05	-7.30	0.000	***
	1	-2.35E-07	3.68E-07	-11.31	0.000	***	2	-2.48E-07	5.18E-06	-0.85	0.397	
	2	-1.20E-07	5.27E-07	-4.02	0.000	***	3	4.14E-07	8.37E-06	0.88	0.382	
	3	-3.78E-08	2.49E-07	-2.68	0.008	***	4	7.42E-07	6.83E-06	1.92	0.056	*
	4	-6.21E-09	1.27E-07	-0.87	0.387		5	8.58E-07	6.20E-06	2.45	0.015	**
	5	-1.24E-08	2.04E-07	-1.08	0.282		6	8.43E-07	4.07E-06	3.66	0.000	***
	6	-3.37E-08	3.84E-07	-1.55	0.122		7	6.46E-07	4.62E-06	2.48	0.014	**
	7	3.08E-08	6.43E-07	0.85	0.397		8	1.74E-07	4.22E-06	0.73	0.468	
	8	-1.20E-08	3.10E-07	-0.69	0.494		9	1.27E-06	1.01E-05	2.22	0.027	**
	9	2.14E-09	1.78E-07	0.21	0.831		10	5.86E-07	3.11E-06	1.10	0.272	
10	8.44E-09	1.47E-07	1.02	0.310								

**Table A.22:** Adaptive Lasso Coefficients LG Display (034220 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.116	0.060	-34.05	0.000	***	1	13.366	4.485	52.72	0.000	***
	2	-0.060	0.038	-28.41	0.000	***	2	-20.311	6.860	-52.38	0.000	***
	3	-0.037	0.032	-20.62	0.000	***	3	-9.767	4.326	-39.95	0.000	***
	4	-0.024	0.025	-16.59	0.000	***	4	-7.624	3.744	-36.02	0.000	***
	5	-0.015	0.023	-11.25	0.000	***	5	-5.113	3.080	-29.37	0.000	***
	6	-0.011	0.021	-9.02	0.000	***	6	-3.557	2.536	-24.81	0.000	***
	7	-0.006	0.016	-6.85	0.000	***	7	-2.227	1.911	-20.62	0.000	***
	8	-0.004	0.022	-3.29	0.001	***	8	-1.270	1.474	-15.24	0.000	***
	9	-0.002	0.009	-3.59	0.000	***	9	-0.680	0.998	-12.05	0.000	***
	10	-0.002	0.012	-2.95	0.003	***	10	-0.284	0.659	-7.64	0.000	***
Trades	0	9.69E-04	7.90E-04	21.69	0.000	***	1	0.275	0.081	60.05	0.000	***
	1	1.66E-04	1.59E-04	18.43	0.000	***	2	0.071	0.052	24.23	0.000	***
	2	1.58E-04	1.47E-04	19.04	0.000	***	3	0.078	0.016	86.10	0.000	***
	3	6.14E-05	8.74E-05	12.43	0.000	***	4	0.055	0.015	64.03	0.000	***
	4	5.48E-05	1.04E-04	9.34	0.000	***	5	0.047	0.013	62.13	0.000	***
	5	2.78E-05	6.93E-05	7.09	0.000	***	6	0.038	0.014	47.93	0.000	***
	6	1.86E-05	4.72E-05	6.99	0.000	***	7	0.033	0.013	47.32	0.000	***
	7	1.12E-05	3.25E-05	6.09	0.000	***	8	0.030	0.012	43.69	0.000	***
	8	7.44E-06	3.40E-05	3.87	0.000	***	9	0.028	0.013	37.91	0.000	***
	9	6.33E-06	3.18E-05	3.53	0.000	***	10	0.031	0.012	45.53	0.000	***
Durations	0	4.12E-06	3.72E-06	19.61	0.000	***	1	-5.80E-06	1.10E-04	-0.94	0.350	
	1	1.58E-07	3.11E-06	0.90	0.369		2	6.57E-06	1.14E-04	1.02	0.308	
	2	7.13E-08	4.77E-06	0.26	0.791		3	1.07E-05	1.05E-04	1.80	0.073	*
	3	1.58E-07	4.60E-06	0.61	0.543		4	2.61E-06	8.96E-05	0.52	0.607	
	4	1.18E-07	3.35E-06	0.63	0.532		5	8.62E-06	7.92E-05	1.93	0.055	*
	5	-2.55E-07	2.70E-06	-1.67	0.096	*	6	7.11E-06	6.47E-05	1.94	0.053	*
	6	-1.23E-07	3.97E-06	-0.55	0.585		7	1.03E-06	6.44E-05	0.28	0.778	
	7	7.24E-08	2.54E-06	0.50	0.614		8	-1.76E-07	7.54E-05	-0.04	0.967	
	8	-3.39E-08	2.16E-06	-0.28	0.781		9	2.13E-06	4.61E-05	0.82	0.416	
	9	7.78E-08	2.20E-06	0.63	0.531		10	2.41E-07	5.23E-05	0.08	0.935	
Spreads	0	1.08E-04	2.54E-04	7.51	0.000	***	1	1.26E-03	6.36E-03	3.52	0.000	***
	1	4.83E-05	2.05E-04	4.16	0.000	***	2	4.09E-04	3.74E-03	1.93	0.054	*
	2	2.59E-05	1.63E-04	2.82	0.005	***	3	2.89E-04	3.00E-03	1.70	0.090	*
	3	1.44E-05	1.23E-04	2.07	0.039	**	4	3.15E-04	2.55E-03	2.19	0.029	**
	4	3.73E-06	8.96E-05	0.74	0.462		5	1.93E-04	2.39E-03	1.43	0.153	
	5	1.17E-05	1.10E-04	1.87	0.062	*	6	2.77E-04	1.61E-03	3.05	0.002	***
	6	5.11E-06	7.79E-05	1.16	0.247		7	1.92E-04	1.97E-03	1.73	0.085	*
	7	-4.78E-06	7.46E-05	-1.13	0.258		8	1.51E-04	1.83E-03	1.46	0.146	
	8	7.30E-06	8.44E-05	1.53	0.127		9	-3.35E-05	1.15E-03	-0.51	0.608	
	9	-8.59E-06	7.51E-05	-2.02	0.044	**	10	9.54E-05	2.57E-03	0.66	0.512	
Depth	0	-3.51E-07	3.73E-07	-16.66	0.000	***	1	-1.19E-05	8.63E-06	-24.34	0.000	***
	1	-4.67E-08	6.62E-08	-12.49	0.000	***	2	-1.04E-06	4.18E-06	-4.39	0.000	***
	2	-2.09E-08	5.11E-08	-7.24	0.000	***	3	-2.83E-07	1.97E-06	-2.54	0.012	**
	3	-8.57E-09	3.23E-08	-4.69	0.000	***	4	3.53E-07	2.63E-06	2.38	0.018	**
	4	-3.66E-09	4.17E-08	-1.55	0.121		5	3.49E-07	1.46E-06	4.22	0.000	***
	5	3.28E-11	1.84E-08	0.03	0.975		6	7.64E-07	2.78E-06	4.87	0.000	***
	6	6.92E-09	1.21E-07	1.01	0.312		7	8.48E-07	2.62E-06	5.73	0.000	***
	7	6.73E-10	1.74E-08	0.68	0.494		8	1.29E-06	3.09E-06	7.41	0.000	***
	8	3.03E-09	1.47E-08	3.64	0.000	***	9	1.37E-06	3.18E-06	7.62	0.000	***
	9	2.94E-09	2.78E-08	1.88	0.062	*	10	2.03E-06	3.49E-06	1.10	0.272	
10	4.62E-09	2.14E-08	3.82	0.000	***							

## A. SUPPLEMENT FOR CHAPTER 4

**Table A.23:** Adaptive Lasso Coefficients Kangwon Land Inc (035250 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.163	0.056	-51.40	0.000	***	1	9.729	2.975	57.87	0.000	***
	2	-0.064	0.043	-26.15	0.000	***	2	-12.497	3.986	-55.47	0.000	***
	3	-0.031	0.030	-18.58	0.000	***	3	-3.120	2.102	-26.26	0.000	***
	4	-0.018	0.027	-12.18	0.000	***	4	-1.772	1.582	-19.81	0.000	***
	5	-0.010	0.024	-7.83	0.000	***	5	-0.561	0.951	-10.43	0.000	***
	6	-0.006	0.015	-6.95	0.000	***	6	-0.267	0.594	-7.95	0.000	***
	7	-0.004	0.013	-5.08	0.000	***	7	-0.155	0.445	-6.15	0.000	***
	8	-0.003	0.010	-4.68	0.000	***	8	-0.051	0.240	-3.75	0.000	***
	9	-0.001	0.011	-2.14	0.033	**	9	-0.025	0.155	-2.82	0.005	***
	10	-0.002	0.010	-3.40	0.001	***	10	-0.011	0.133	-1.46	0.146	
Trades	0	2.42E-03	1.40E-03	30.62	0.000	***	1	0.358	0.096	65.81	0.000	***
	1	3.46E-04	3.84E-04	15.94	0.000	***	2	0.075	0.053	25.02	0.000	***
	2	2.10E-04	2.65E-04	14.05	0.000	***	3	0.071	0.023	54.72	0.000	***
	3	8.68E-05	1.92E-04	7.99	0.000	***	4	0.036	0.021	30.25	0.000	***
	4	5.35E-05	1.47E-04	6.43	0.000	***	5	0.029	0.018	28.34	0.000	***
	5	1.77E-05	6.07E-05	5.15	0.000	***	6	0.022	0.018	21.13	0.000	***
	6	7.59E-06	5.96E-05	2.25	0.025	**	7	0.017	0.015	20.02	0.000	***
	7	2.70E-06	4.48E-05	1.06	0.288		8	0.016	0.015	18.11	0.000	***
	8	4.04E-06	4.91E-05	1.46	0.147		9	0.015	0.015	17.65	0.000	***
	9	2.26E-07	4.15E-05	0.10	0.923		10	0.016	0.016	18.51	0.000	***
10	-3.23E-06	4.01E-05	-1.42	0.155								
Durations	0	6.64E-06	5.72E-06	20.56	0.000	***	1	-2.62E-06	1.08E-04	-0.43	0.667	
	1	3.63E-07	4.61E-06	1.39	0.165		2	-6.24E-06	6.66E-05	-1.66	0.098	*
	2	7.09E-08	2.58E-06	0.49	0.628		3	1.68E-06	4.30E-05	0.69	0.491	
	3	-5.29E-08	2.70E-06	-0.35	0.729		4	8.41E-07	3.33E-05	0.45	0.656	
	4	-3.31E-07	2.98E-06	-1.97	0.050	*	5	4.41E-07	4.04E-05	0.19	0.847	
	5	2.55E-07	2.36E-06	1.92	0.056	*	6	-2.87E-06	2.97E-05	-1.71	0.089	*
	6	-4.49E-08	1.57E-06	-0.51	0.613		7	1.11E-06	1.75E-05	1.12	0.262	
	7	-9.56E-08	2.15E-06	-0.79	0.431		8	-1.39E-06	1.09E-05	-2.24	0.026	**
	8	-1.20E-07	1.48E-06	-1.43	0.153		9	3.48E-07	3.16E-05	0.19	0.846	
	9	-1.98E-07	1.92E-06	-1.83	0.069	*	10	-4.22E-07	9.21E-06	-0.81	0.418	
10	1.64E-07	1.89E-06	1.54	0.125								
Spreads	0	1.51E-04	3.22E-04	8.31	0.000	***	1	8.36E-04	4.78E-03	3.10	0.002	***
	1	3.10E-05	2.43E-04	2.26	0.025	**	2	4.18E-04	3.81E-03	1.94	0.053	*
	2	1.71E-05	1.55E-04	1.95	0.052	*	3	7.77E-05	2.16E-03	0.64	0.525	
	3	1.28E-05	1.94E-04	1.16	0.246		4	-1.53E-04	2.28E-03	-1.19	0.236	
	4	1.94E-05	1.50E-04	2.28	0.023	**	5	-1.98E-05	1.53E-03	-0.23	0.819	
	5	-1.96E-06	9.82E-05	-0.35	0.725		6	3.22E-05	1.16E-03	0.49	0.623	
	6	2.74E-06	1.01E-04	0.48	0.632		7	-1.58E-05	5.76E-04	-0.49	0.627	
	7	-7.03E-06	1.21E-04	-1.03	0.304		8	-4.44E-06	6.47E-04	-0.12	0.904	
	8	-5.55E-06	9.22E-05	-1.07	0.287		9	-4.71E-05	8.14E-04	-1.02	0.307	
	9	2.98E-06	6.49E-05	0.81	0.417		10	7.92E-06	6.46E-04	0.22	0.828	
10	6.97E-07	8.03E-05	0.15	0.878								
Depth	0	-1.48E-06	1.14E-06	-22.96	0.000	***	1	-1.33E-05	2.11E-05	-11.20	0.000	***
	1	-2.09E-07	3.45E-07	-10.71	0.000	***	2	-9.04E-07	7.48E-06	-2.14	0.033	**
	2	-4.94E-08	2.16E-07	-4.04	0.000	***	3	-1.76E-07	3.01E-06	-1.04	0.301	
	3	-1.91E-08	2.25E-07	-1.50	0.135		4	3.89E-07	3.33E-06	2.06	0.040	**
	4	-1.34E-08	1.88E-07	-1.26	0.210		5	2.58E-07	2.55E-06	1.79	0.075	*
	5	-2.89E-09	8.43E-08	-0.61	0.545		6	4.85E-07	2.77E-06	3.10	0.002	***
	6	3.48E-09	7.87E-08	0.78	0.434		7	7.49E-07	4.90E-06	2.70	0.007	***
	7	1.98E-08	1.31E-07	2.69	0.008	***	8	1.08E-06	5.01E-06	3.82	0.000	***
	8	7.53E-09	2.15E-07	0.62	0.535		9	5.62E-07	2.83E-06	3.51	0.001	***
	9	6.12E-09	1.67E-07	0.65	0.517		10	5.90E-07	3.40E-06	1.10	0.272	
10	2.24E-08	1.13E-07	3.52	0.000	***							

**Table A.24:** Adaptive Lasso Coefficients LG Household & Health Care (051900 KS)

We estimate the coefficients of the adaptive lasso VARX model using high frequency tick data on a weekly basis from Jan 2007 to Dec 2012. Here we show the average coefficients across time.

	Quote Revision Equation					Trade Equation						
	lag	mean	stdev	T-stat	p-value	lag	mean	stdev	T-stat	p-value		
Quote Revisions	1	-0.163	0.060	-48.41	0.000	***	1	6.258	2.252	49.16	0.000	***
	2	-0.056	0.037	-26.68	0.000	***	2	-3.987	1.522	-46.35	0.000	***
	3	-0.028	0.033	-14.96	0.000	***	3	-0.657	0.830	-14.01	0.000	***
	4	-0.012	0.020	-10.81	0.000	***	4	-0.235	0.454	-9.17	0.000	***
	5	-0.005	0.017	-5.47	0.000	***	5	-0.050	0.234	-3.80	0.000	***
	6	-0.004	0.014	-5.34	0.000	***	6	-0.034	0.182	-3.30	0.001	***
	7	-0.002	0.010	-3.01	0.003	***	7	-0.008	0.060	-2.44	0.015	**
	8	-0.001	0.008	-3.15	0.002	***	8	-0.010	0.079	-2.34	0.020	**
	9	-0.001	0.007	-3.16	0.002	***	9	-0.002	0.039	-1.12	0.264	*
	10	0.000	0.005	-0.88	0.379		10	-0.002	0.025	-1.71	0.088	*
Trades	0	5.75E-03	3.06E-03	33.19	0.000	***	1	0.302	0.110	48.45	0.000	***
	1	1.14E-03	9.02E-04	22.32	0.000	***	2	0.070	0.040	31.16	0.000	***
	2	6.16E-04	6.97E-04	15.63	0.000	***	3	0.055	0.022	44.13	0.000	***
	3	2.35E-04	4.02E-04	10.33	0.000	***	4	0.032	0.019	28.92	0.000	***
	4	8.75E-05	2.52E-04	6.14	0.000	***	5	0.024	0.018	22.83	0.000	***
	5	6.75E-05	2.30E-04	5.18	0.000	***	6	0.018	0.018	18.28	0.000	***
	6	1.88E-05	1.41E-04	2.36	0.019	**	7	0.014	0.015	16.70	0.000	***
	7	3.02E-06	4.28E-05	1.25	0.213		8	0.012	0.013	15.70	0.000	***
	8	-4.21E-06	1.16E-04	-0.64	0.521		9	0.012	0.015	14.59	0.000	***
	9	-1.05E-05	1.24E-04	-1.49	0.136		10	0.012	0.014	15.50	0.000	***
10	-2.44E-05	1.45E-04	-2.99	0.003	***							
Durations	0	9.86E-06	6.88E-06	25.37	0.000	***	1	-1.77E-06	3.29E-05	-0.95	0.343	
	1	-3.96E-09	1.38E-06	-0.05	0.960		2	-1.42E-06	1.99E-05	-1.26	0.207	
	2	-6.55E-08	1.06E-06	-1.10	0.274		3	3.07E-07	6.40E-06	0.85	0.397	
	3	4.65E-08	1.51E-06	0.55	0.586		4	2.11E-08	3.24E-06	0.12	0.908	
	4	1.16E-09	1.08E-06	0.02	0.985		5	-6.19E-07	9.15E-06	-1.20	0.232	
	5	-1.46E-07	3.69E-06	-0.70	0.484		6	6.90E-08	4.90E-06	0.25	0.804	
	6	-1.61E-08	8.87E-07	-0.32	0.748		7	5.20E-08	1.37E-06	0.67	0.503	
	7	3.44E-07	5.96E-06	1.02	0.307		8	-2.03E-07	3.00E-06	-1.20	0.231	
	8	-1.25E-07	1.43E-06	-1.54	0.124		9	3.59E-08	2.07E-06	0.31	0.759	
	9	-8.12E-08	1.18E-06	-1.21	0.226		10	-1.01E-07	1.96E-06	-0.92	0.360	
10	-6.75E-08	1.02E-06	-1.17	0.243								
Spreads	0	7.72E-06	1.41E-05	9.71	0.000	***	1	1.61E-05	7.76E-05	3.68	0.000	***
	1	2.75E-06	1.01E-05	4.83	0.000	***	2	5.16E-06	5.09E-05	1.79	0.074	*
	2	1.34E-06	9.66E-06	2.45	0.015	**	3	-1.95E-06	2.21E-05	-1.56	0.119	
	3	1.58E-07	6.64E-06	0.42	0.673		4	-2.10E-07	2.70E-05	-0.14	0.891	
	4	9.51E-07	7.61E-06	2.21	0.028	**	5	1.32E-06	3.95E-05	0.59	0.555	
	5	2.03E-07	5.19E-06	0.69	0.490		6	-2.21E-07	1.31E-05	-0.30	0.765	
	6	-4.54E-07	5.21E-06	-1.54	0.124		7	-1.25E-07	7.79E-06	-0.28	0.776	
	7	-2.06E-07	5.50E-06	-0.66	0.508		8	7.04E-07	1.31E-05	0.95	0.343	
	8	-1.00E-06	7.60E-06	-2.33	0.020	**	9	1.64E-06	2.79E-05	1.04	0.299	
	9	-3.02E-07	6.22E-06	-0.86	0.391		10	-4.08E-08	1.74E-05	-0.04	0.967	
10	-7.71E-07	6.78E-06	-2.01	0.045	**							
Depth	0	-1.87E-05	1.55E-05	-21.29	0.000	***	1	1.54E-05	9.17E-05	2.97	0.003	***
	1	-3.52E-06	7.76E-06	-8.03	0.000	***	2	3.83E-06	4.69E-05	1.44	0.150	
	2	-1.71E-06	6.21E-06	-4.88	0.000	***	3	7.98E-06	4.83E-05	2.93	0.004	***
	3	-2.18E-07	5.36E-06	-0.72	0.473		4	5.53E-06	3.86E-05	2.53	0.012	**
	4	-4.41E-07	3.99E-06	-1.96	0.051	*	5	5.48E-06	3.93E-05	2.47	0.014	**
	5	-1.50E-07	2.84E-06	-0.93	0.351		6	5.34E-06	3.46E-05	2.73	0.007	***
	6	7.67E-08	2.08E-06	0.65	0.515		7	5.01E-06	3.13E-05	2.83	0.005	***
	7	4.55E-09	2.17E-06	0.04	0.970		8	3.71E-06	3.19E-05	2.06	0.040	**
	8	1.21E-07	2.28E-06	0.94	0.349		9	5.36E-06	2.98E-05	3.18	0.002	***
	9	-1.47E-07	1.92E-06	-1.36	0.175		10	1.88E-06	1.42E-05	1.10	0.272	
10	-1.01E-07	1.98E-06	-0.90	0.368								