MONASH University

Stock Return Prediction with Hidden Order Mapping

Varsha Mamidi

A thesis submitted for the degree of Doctor of Philosophy
at
Monash University in 2016

Faculty of Information Technology

Stock Return Prediction with Hidden Order Mapping

ABSTRACT

Missing data problem is ubiquitous in many real life situations. Information Technology researchers have explored and tried to address this problem in different settings. In this thesis, we undertake research to address missing data problem associated with order book information in stock markets. This is an in-depth and large-scale study with systematic and comprehensive framework to address missing data problem in the finance literature.

Orders placed by traders and the corresponding order imbalance (OIB) is informative to predict future stock returns, however, stock exchange rules do not reveal price sensitive complete order book data for traders. Hence, return prediction using the revealed, incomplete trade book data (that contains only matched buy and sell orders and deletes the unmatched orders), does not let traders to completely exploit possible short term trading opportunities. Hence, this can be considered as a classical missing data problem for predicting future returns, by using the information content of order book. This thesis addresses the missing data problem by developing an integrated theoretical framework applied in stock market trading environment. We use relational Markov

networks theory and build an empirically testable Algorithm for Imputed Complete Order Book (AICOB).

The thesis contributes by developing a new theoretical advancement of information technology research relating to missing data problem and applying it to financial markets. First, the thesis presents the missing data problem as a Missing at Random (MAR) data and builds a systematic framework to estimate single as well as joint log likelihood functions. The thesis demonstrates that estimating by using incomplete records, improves the accuracy of the parameter estimates.

Second, the thesis proposes a Relational Markov Network Model for estimation of the joint distribution function of orders, order characteristics and their interactions. Later, the Expectation Maximization Algorithm is proposed to address the missing data problem during the joint estimation procedure. All pooled regression results follow Fama and MacBeth (1973) and Generalized Methods of Moments (GMM) methods. These methods control the cross sectional and time series correlations between the observations and across the pooled stocks. The proposed novel methodology overcomes the estimation problem in the context of missing order book data.

Third, the thesis develops an objective evaluation strategy for AICOB, based on efficiency, accuracy and adaptability dimensions. The thesis uses Australian stock market data, which provides not only trade book data but also historical order book data for cross validating the results. This unique setting allows validating the accuracy of AICOB methodology by comparing with complete order book data. The main

contribution of the thesis is to show that AICOB based predictions match with the complete order book data. Whereas, trade book based predictions are quite inconsistent to the complete order book data. The AICOB based results are also consistent with the theoretical predictions proposed in finance literature that OIB predicts better as the firm size increases. The results show that large firms, with higher trading activity and more competition for order flows, report more significant OIB prediction of future returns. Trade Book based OIB estimates, which suffer from missing data problem, fail to predict future returns for stock portfolios. Hence, addressing the missing data problem is important before implementing OIB based trading strategies.

Overall, the thesis finds that machine learning applications, similar to AICOB, can be helpful in implementing the trading strategies in financial markets. The imputation of missing data, through systematic procedures, based on theoretical distributional properties of the financial variables, can be informative for more accurate prediction of future returns. The thesis contributes towards establishing a common ground for cross-disciplinary research in Finance and Information Technology (IT) by applying the advances in IT research to solve research and corresponding implementation problems in finance. Further, the thesis contributes towards advancing the order imbalance literature by showing that missing data can play a critical role in the predictive ability of order imbalance.

Stock Return Prediction with Hidden Order Mapping

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Varsha Mamidi

August 3, 2016

Dedicated to

the forces of the universe…

in which the energies of animate and inanimate dwell…

that fueled and propelled my engines of learning…

LIST OF PUBLICATIONS

Publications:

- Kumar, K., Mamidi, V., & Marisetty, V. (2011). Global markets exposure and price efficiency: An empirical analysis of order flow dynamics of NYSE-listed Indian firms. *Journal of International Financial Markets, Institutions & Money*, 21(5), 686-706. (PhD proposal-based paper).

Awards:

- Best paper award at the third doctoral colloquium organised by Institute of Development Banking Research and Technology, India, 2013, http://www.idrbt.ac.in/news/idc2013.html, last accessed on 31/07/2016.

ACKNOWLEDGMENTS

Foremost, I would like to extend my gratitude to my supervisors: Prof. Bala Srinivasan, Dr. Huu Duong and Prof. Madhu Veeraraghavan for guiding me in the PhD journey. This thesis would not have seen the light of the day, if it were not for their tremendous help and support. I sincerely thank Madhu, for encouraging me to take up PhD studies. I would not have embarked on this journey without his support. I convey my deepest gratitude to Srini, who has relentlessly helped and supported me throughout the journey. I gratefully thank him for his insight and guidance at every stage of the thesis. I sincerely thank Huu for his continuous enthusiasm and guidance. I also thank him for uplifting my spirits at difficult times that kept me going.

I would like to thank Dr. Kiran Kumar Kotha for his initial encouragement and help in understanding nuances of the stock market data. I would also like to express my special thanks to Dr. Yogesh Chouhan for his help and guidance with SAS, during the early stages.

I would like to extend my special thanks to Allison for always being there to help me. She always lightened the stress with her encouraging words and a positive outlook. I would also like to thank Michelle Ketchen, Diana Sussman, Ra_g Tjahjadi, Denyse Cove, and Akamon Kunkongkapun for always being helpful with a cheerful smile and providing me with administrative and technical support during my research.

I would also like to thank Prof. Frada Burstein, Dr. Michael Morgan, Dr. Mark Carman, Dr. Grace Rumantir, Dr. Maria Indrawan and Dr.Paul Lajbcgier for their advice and assistance in numerous capacities.

I would like to express my gratitude to the jury members of the doctorial colloquium at IDRBT and the research fellows at various conferences for their suggestions, ideas and advice.

I would also like to extend my appreciation to all my colleagues and the fellow PhD students. I would like to express special thanks to Dwi Rahayu for her constant support throughout the journey. I would also like to thank Abdullah M. Almuhaideb, Sepehr Minagar, Zahraa Abdallah, Peter Serwylo, Mark Creado, Danny Ardianto, Shan He (Dora), Ram Kumar from IIT Bombay and many other friends at Caulfield School of IT who patiently sat through many mock presentations, supported and helped me through this entire process and made the candidature period enjoyable.

I would like to express my deepest gratitude to my parents, my close family members and friends for their support and prayers. Their unconditional love and continued blessings helped me accomplish my dream.

Last but not the least; I am very grateful to my dear husband and my darling daughter for their continuous love and support. I thank my husband from the bottom of my heart for being the source of my energy, for being my friend, philosopher and guide; who encouraged, helped and supported me in this long tiresome journey. I thank him for never leaving my side, for brightening my day with his ideas, for improving my critical

thinking with his questioning, for pushing me when I felt lazy, for picking me up when I felt broken, for being the beacon of light when I felt lost. I thank my daughter for all the love and patience. She will be the most happiest to know that it is completed, *at last*.

Finally, I thank all my well-wishers and my guardian angels for their love, support, patience and prayers.

Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Preamble

Stock markets are major platforms for capital creation, valuation, and investments. With a stock valuation of global stocks more than USD $118 trillion, the global stock market is one of the largest and significant service sector in the world.[1] The most critical element of a well-functioning financial market is whether assets are being traded at fair prices and how market mechanism facilitates their price discovery process (price formation process).

A market is a place where goods and services are bought and sold. Buyers and sellers can be individuals, firms, factories, agents or dealers. In the case of an equities stock market, financial contracts known as shares (or stocks), of listed firms are traded by buyers and sellers through their brokers or market agents. The rules, policies and laws that govern the market, in which the buyers and sellers interact with each other to determine the price and quantity of the goods and services to be exchanged, is termed as market mechanism. A clearing price is arrived when the demand schedule for a particular quantity ordered matches with the supply schedule at the same price and quantity. This is normally termed as the equilibrium price.

[1] Source: World Federation of Exchanges 2013 report. The value includes, equities, derivatives, bonds, exchange traded funds and securitized derivative.

Price discovery process involves traders' arriving at a new equilibrium price for a given stock in a competitive trading environment. This is a dynamic process where stock price direction changes due to changes in the information content relating to firms or traders or general market related activity and corresponding temporal inventory imbalances of stocks (demand or supply imbalances). Market mechanism and trading rules play a significant role in market efficiency. A trading rule, for instance, imposing a limit on the price at which a stock is bought or sold is very common in many markets. Stock exchanges impose such rules to avoid possible market crashes. For instance, during market panic, traders engage in pushing the price below its natural equilibrium due to their flight for liquidity.[2] Hence, if a stock exchange mandates traders that they cannot trade a stock below a certain price on a given trading day, they can avoid sudden drop in the stock price that may trigger market crash. However, such rule imposes the stock not be traded at its true equilibrium. Hence, the true price cannot be discovered on that day due to trading mechanism related rule. Thus, such rules and regulations can play an important role in the price discovery process. Any delay in the price discovery process gives rise to market inefficiency.

1.1.1 Market Efficiency

The speed at which information gets incorporated into asset prices (Fama, 1970) determines the efficiency of the markets. Hence, the delay in the price discovery process not only reflects market inefficiency, but also provides opportunities to engage in profitable arbitrage trades. This implies that any inefficient price changes are bound to

[2] Flight for liquidity: Traders liquidating their position to quickly leave the market.

go back to their original values. Hence, one can engage in profitable arbitrage by conducting offsetting trades. For instance, if market price goes up artificially then it is bound to come down to its equilibrium price at a later time. In such an instance, an arbitrageur can sell the stock when price goes up and offset his/her trade position by taking an opposite buy position when price comes down. The positive difference due to buying at low price and selling at high price is the arbitrageur's trading profit. Market can be inefficient due to less transparency of orders and temporal nature of order imbalance. Orders, when hidden, do not reveal their influence (to the traders) on the trade direction and corresponding share price changes. Hence, there is a higher chance that traders can misprice their orders without knowing the actual market demand. Traders around the world, in order to exploit arbitrage opportunities, use models and techniques to infer trade direction and the corresponding price adjustment process (Hendershott et al., 2011). However, such short-term inventory based arbitrage opportunities are hard to consistently exploit due to complex, incomplete and voluminous data.

1.1.2 Market Mechanism

Orders submitted by investors are mainly categorised as buy and sell orders. Buy orders are referred to as *bids* and sell orders are referred to as *asks*. Figure 1.1 depicts the general framework of an electronic stock market. Orders are made through stock brokers' electronic terminals (information systems) which are later processed by order routing systems into an execution system. Once the 'bids' and 'asks' are matched through the trading execution system, the traded orders move into clearing system to complete the transaction.

The execution system maintains two main books. One that records all orders (order book) and the other that records only the matched orders (trade book). The order book in electronic markets is also termed as *limit order book* as investors are supposed to give a price while placing an order.

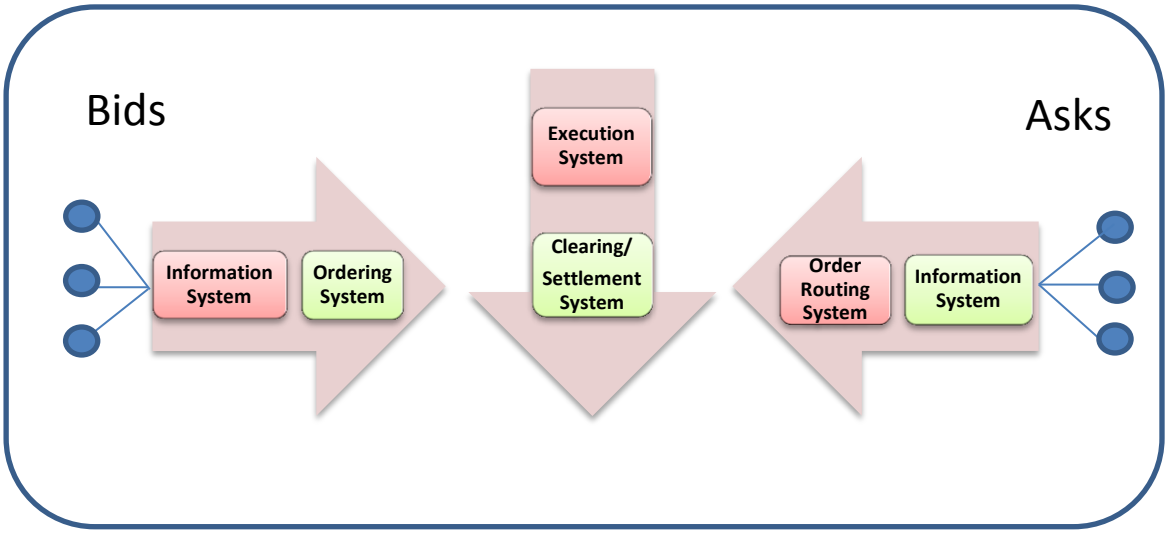


Figure 1.1 Electronic Market Mechanism

1.1.3 Snapshot of an Order Book

A snapshot of limit order book of Microsoft stock at the NASDAQ stock exchange, as on 24th May 2010 is depicted in Figure 1.2. As shown in this figure, there could be hundreds of orders on both the buy side and the sell side at a given time stamp (in this case it is 13.15.23.745 hours). Also, the orders can be colour coded indicating different types of orders like buy orders and sell orders. At a given point of time, there may be an imbalance (buys more than sells or sells more than buys) between buy and sell orders. For instance, buy orders might be more than sell orders. Such an *order imbalance* (OIB) indicates that there are more buyers compared to sellers. This could imply that the

possible future trade direction or price change is likely to be positive. Similarly, seller initiated order imbalance could lead to negative price change. However, acting on this information can be deceptive if the stock exchanges censor the complete order book data. Stock exchanges mainly reveal only those orders that are matched and resulting into a trade. Such matched orders are displayed in a separate book called trade book.

Figure 1.2 Microsoft Limit Order Book as on 24th May 2010 at 13.15 hrs

(Source: NASDAQ Stock Exchange)

The observable data in the trade book can only provide traded order imbalance which is useful but cannot be complete for predictions because traded order imbalance contains only matched orders data. It will not contain all the orders that are part of the complete supply and demand schedule at that point of time. It contains only the order pairs that are matched. The value of such order imbalance information content is very

short lived due to frequent arrival or updating of orders. Therefore, OIB will not persist over time, leading to auto regressive (AR) process. This short interval constraint challenges traders to profit from such temporal order imbalance. Hence, this study aims at devising a framework to address missing orders data problem by integrating both information systems and finance disciplines. From information systems research point of view, understanding and modeling missing data in a complex, dynamic and big data setting, like stock market is an important research issue. Likewise, from finance research point of view, understanding the influence and implications of complete order book on price discovery process and market efficiency is an unresolved issue. This cross-disciplinary complementarity serves as the source of motivation for this thesis.

1.2 Stock Market Behaviour

With improvements in computing ability and advancements in machine intelligence research, information systems and machine learning algorithms have become a significant part of modern day trading. Hendershott et al. (2011) report that stock market trading that uses computer based algorithms constitute 73% of the total trading volume in the U.S. market. They also find that algorithm based trading, that involves information generation from visible orders and placement of orders based on some pre-set trading rules, has improved overall market quality. Hence, applied research in information systems area that tries to improve the accuracy of stock returns prediction and understanding the possible impact of missing data in a complex environment like stock market, would not only have the potential to uncover profitable trading strategies, but also improves overall market welfare by increasing market efficiency through faster price discovery process.

However, existing applied research on the efficacy of information systems, in terms of addressing a complex finance problem, is still at its nascent stage. This could be due to non-availability of voluminous data on stock trading for research purposes. Although stock market data, at daily intervals, have been available for a few decades, the rich voluminous intra-day data are made available only in the recent past. In this thesis, research based on machine learning methods is applied to address this specific challenge that arises due to trading rules set by stock exchanges for censoring data.

Also, existing literature measures order imbalance by using trade book data,[3] as the complete order book is not visible to traders. Researchers have to rely on some assumptions to estimate the order imbalance. For instance, finance studies use the Lee and Ready (1991) algorithm to assign the traded orders into buyer initiated orders and seller initiated orders (Chordia et.al, 2005). In the Lee-Ready algorithm, a trade is classified as buyer/seller initiated if it is closer to bid/ask of the prevailing quote. The quote must be at least five seconds old. If the trade is exactly at the midpoint of the quote, a "tick test" is used whereby the trade is classified as buyer/seller initiated if the last price change prior to the trade is positive/negative. However, the Lee and Ready algorithm is not a realistic algorithm in algorithmic trading context as it ignores trades that occur less than 5 seconds from the current trade. Also, recent research has found that the Lee and Ready algorithm systematically misclassifies trades at the mid-point bid-ask spread and thus leads to inaccurate trade classification (White, 2000). Using a more systematic framework that infers order book data and imputes missing information

[3] Please refer to chapter 2 for more information on Trade and Order books.

of order book has the potential to improve the estimated level of the order imbalance and therefore helps in better prediction of future returns.

1.2.1 Importance of Order Imbalance on Future Price: Walmart Example

To get some perspective on the role of order imbalance on future price movements or trade direction, we report Walmart stock price movement for one day (September 30, 2010) in the Figure 1.3. There are two panels reported in this figure. Top panel, Panel 1, shows price movement and the bottom panel, Panel 2, shows order imbalance. In Panel 1, volume traded during the day is represented through blue vertical line bars; price movement is represented by red line graph for that particular day. We also report major news associated with Walmart stock announcements in purple line bars in Panel 1. In Panel 2, the red line bars, represent order imbalance for the corresponding timestamps which are synchronized with both the panels. This means that the price at any given time in Panel 1 corresponds to the order imbalance for the same time in Panel 2. The red line bars above the origin (in Panel 2) indicate positive imbalance (more buy orders compared to sell orders) and below the origin lines indicate negative imbalance (more sell orders than buy orders). Panel 2 shows positive imbalance for the first part of the day (before 11:30 am). This causes the price in Panel 1 to go up after 11:50 am (approximately). This demonstration of such leading influence of order imbalance on price changes is the main premise on which this thesis is built.

1.2.2 Volume vs Order Imbalance

Order imbalance has a permanent effect on price formation (Chordia et al., 2000). It can predict future prices better than volume. Volume has only magnitude where as OIB has

both magnitude and direction. As shown in the Figure 1.3, volume lines in the Panel 1 show more/less volume but that does not impact price change.

The volume movements throughout the data are quite random (ups and downs), however, the price movements have a clear downward trend for couple of hours of trade and then a clear upward trend until the end of the day. This patterns corresponds more to the order imbalance data in Panel 2. A huge positive order imbalance in the early hours of trade has led to a clear upward trade direction for the rest of the day. Unlike order imbalance, it is difficult to draw inference using volume as it does not have a direction. High volume does not necessarily mean high buy orders compared to sell orders.

1.2.3 Evidence with Multiple Stocks

In theory, going by the Walmart example in Figure 1.3, stock prices of Walmart for the rest of the day can be predicted after observing a certain threshold level of order imbalance at a given point of time. However, it is hard to generalize this result without conducting a more robust test by using multiple stocks and for multiple periods of time.

By doing this in the next section, we reinforce the fact that order imbalance can predict prices. We investigate in detail the general implications of the importance of order imbalance. We use six listed stocks from the Bombay Stock Exchange that provide transaction level intra-day data. These stocks are, Dr. Reddy's Laboratories, ICICI Bank, Mahanagar Telephone, Satyam Computer Services, Videsh Sanchar Nigam and Wipro. The objective is to understand how lagged order imbalance values (using the Lee and Ready algorithm, 1991) can influence contemporaneous stock return (percentage price change).

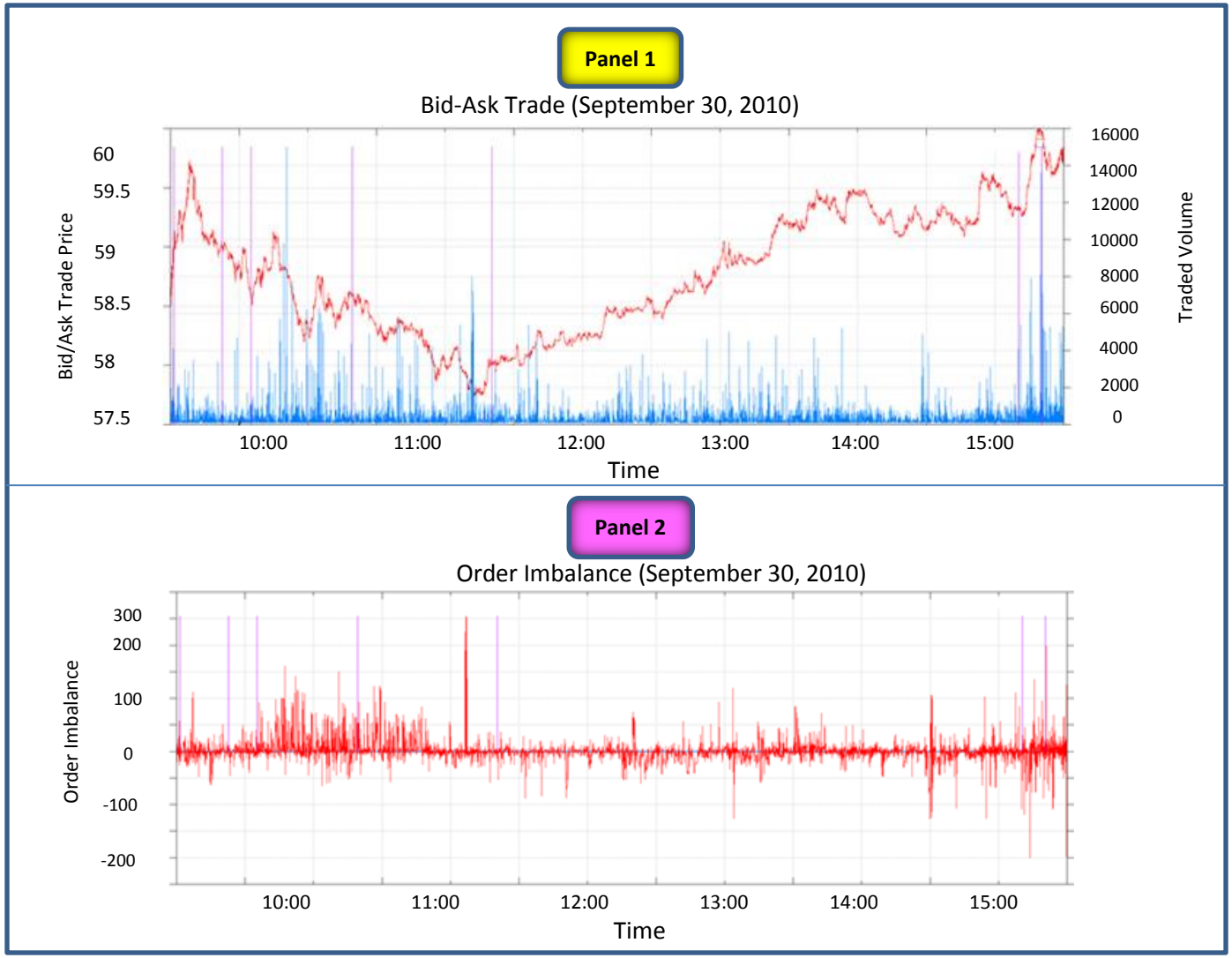


Figure 1.3 Walmart Order Imbalance: An Illustrative Example

Table 1.1 reports correlation between contemporaneous stock return (percentage price change) and lagged order imbalance that is lagged at different time intervals ranging from 5 minutes to one day. As reported in the table, based on the correlation coefficients, contemporaneous return (MR) is significantly positively correlated with lagged order imbalance both in value (OIBV) and numbers (OIBN). This implies that order imbalance can influence future returns.

Table 1.1 The Role of Order Imbalance on Returns (Univariate Analysis)
** represents 5% level of significance

	5 Minutes			***10 Minutes***		
Variable	***MR***	***OIBN***	***OIBV***	***MR***	***OIBN***	***OIBV***
MR	1	0.177**	0.187**	1	0.182**	0.157**
OIBN	0.177**	1	0.375**	0.182**	1	0.355**
OIBV	0.187**	0.375**	1	0.157**	0.355**	1
	15 Minutes			***Daily***		
Variable	***MR***	***OIBN***	***OIBV***	***MR***	***OIBN***	***OIBV***
MR	1	0.182**	0.155**	1	-0.096	0.789
OIBN	0.182**	1	0.342**	-0.096	1	0.072
OIBV	0.155**	0.342**	1	0.789	0.072	1

1.2.4 Evidence with Cross Listed Stocks

We further extend the analysis on the role of order imbalance on price movements. Table 1.2 shows portfolio of six stocks that are cross-listed on the Bombay Stock Exchange, the National Stock Exchange and the New York Stock Exchange. The reason for selecting cross-listed stocks is to ensure that the sample includes stocks that are very efficiently priced (in terms of their speed of adjustment to new information). Listing the same stocks in multiple stock exchanges encourage investors to undertake arbitrage trades to consequently eliminate the possibility of

inefficient pricing. Hence, it is hard[4] to predict future direction of these stock prices. We calculate traded order imbalance (buy orders minus sell orders) from trade book in both number of orders (OIBN) and value of orders (OIBV).

Table 1.2 Multivariate Analysis on the Role of Order Imbalance on Returns

** represents 5% level of significance *** represents 1% level of significance

Model	***R1***	***R2***	***R3***	***R4***	***R5***	***R6***
Type	**60 Minutes**					
MR_{t-1}	-0.385 (7.991)***	-0.385 (8.010)***	-0.387 (8.099)***	-0.388 (8.161)***		
$OIBV_{t-1}$	0.001 (-0.431)	0.001 (-0.52)				0.001 (-1.488)
$OIBN_{t-1}$	0.001 (-0.13)		0.001 (-0.318)		0.004 (-0.898)	
Type	**5 Minutes**					
MR_{t-1}	0.01 (-0.663)	-0.005 (-0.337)	0.007 (-0.464)	-0.005 (-0.319)		
$OIBV_{t-1}$	0.001 (4.643)***	0.001 (-1.032)				0.001 (-1.027)
$OIBN_{t-1}$	0.217 (14.42)***		0.19 (13.66)***		0.189 (13.66)***	

We measure price movements through price changes or returns (MR). The objective is to see whether price changes are predictable by using lagged order imbalance data. We calculate the lags at 5 minutes, 10 minutes, 15 minutes and daily intervals. Table 1.2 reports results based on a regression model where the dependent variable is the contemporaneous return and the explanatory variables are various measures of lagged order imbalance ($OIBV_{t-1}$ and $OIBN_{t-1}$).

[4] Hard to predict as more traders compete in multiple markets so there is no delay in information adjustment in prices.

The regression equation for model R1 is as follows:

$$MR_t = \alpha_0 + \alpha_1 MR_{t-1} + \alpha_2 OIBV_{t-1} + \alpha_3 OIBN_{t-1} + \varepsilon_t \quad (1.1)$$

MR_t represents mid-point return at time *t* of weighted portfolio of six cross listed stocks. MR_{t-1} represents midpoint return of the portfolio during *t-1* period (t = 60 minutes or 5 minutes). $OIBV_{t-1}$ represents the order imbalance value (in Indian Rupees: 1US$ ≈ 65 Rupees) of the portfolio for *t-1* period. $OIBN_{t-1}$ represents the order imbalance in number of orders for the same portfolio. α_0 is the intercept, α_1, α_2, α_3 are the coefficient terms and ε_t is the error term. The values in the parentheses are the t-values and ** represents significance at 5% level. Allowing both $OIBV_{t-1}$ and $OIBN_{t-1}$ in Equation 1.1 also helps to capture the role of nature of trader in the price discovery process. Order imbalance in *number* of orders captures the possibility of a larger number of liquidity motivated traders and order imbalance in *value* captures the role of information traders who act on price sensitive information through a few large trades.

We also include lagged return as a control variable. We report six regression model based results from R1 to R6. Each model is different in terms of the independent variables used. The values in brackets are the t-values. The results in Table 1.2 indicate that order imbalance is positively correlated with returns, however, only at 5 minutes intervals. As the intervals increase to 60 minutes, the correlation disappears. This result highlights two important findings. First, using past order imbalance[5] $OIBN_{t-1}$, one can predict the future price movements as the

[5] Last 5 minutes lagged OIBN for 5 minutes interval; last 60 minutes lagged OIBN for 60 minutes interval.

coefficient for $OIBN_{t-1}$ is significant[6] at 5%. It is denoted with "**" in the three regression models R1, R3 and R5 (bracketed values in the last row of the table 1.2) that are employed to the relationship. For instance, for R1 model, the coefficient of 0.217 indicates that for 100% increase in order imbalance, in the next 5 minutes, prices move by 21.7% in the same direction. It should be noted that the influence of lagged order imbalance on the contemporaneous return can be due to price pressure caused by information traders and also due to the role of liquidity traders who supply liquidity by taking the opposite positions with the information traders.

Second, in order to exploit such potential arbitrage opportunities one should act within 5 minutes. This is because, longer intervals of time, like 60 minutes in the table, indicate that order imbalance does not have any significant influence on the future price direction However, it is relative to the trading activity of a given stock. An infrequently traded stock may need a longer window compared to a frequently traded stock as the price adjustment process of an infrequently traded stock will be much longer. Given that the literature on order imbalance considers a range of time frequency windows from 5 minutes to 60 minutes, the same norm is followed in the thesis. The second finding further strengthens the argument on the usage of machine learning methods to map the order book information.

Table 1.2 reports regression results of the relationship between contemporaneous returns with lagged returns and lagged order imbalance. The results are based on intraday intervals. We report 60 minutes and 5 minutes intervals. Although, we also estimated 10, 15, and 30 minutes interval data, we report only two intervals as the results are almost similar to 60 minutes interval. The other regression

[6] t-value, greater than 1.96.

models ranging from R2 to R6 are different variations of the regression model, R1. It should be noted that the order imbalance used in this study is only the traded order imbalance. The actual imbalance available in the hidden order book would be more informative and hence enhances the predictive ability. This predictive ability of order imbalance provides the basis to undertake research on investigating further on the relationship between order imbalance and future stock returns and thereby addressing research questions that are outlined in the Section 1.4.

1.3 Motivation

Most of the stock exchanges record all transactions in two separate books, namely, order book and trade book. Order book records all the orders (that include fresh orders, modified orders and cancelled orders) placed by investors and trade book includes only a subset of order book data that includes only those orders that are converted into a trade due to matching of buying and selling orders. In other words, order book records both intention to trade and actual trades, however, trade book records only actual trades. In majority of the stock exchanges around the world, trade book with a delay of few minutes is the only visible information to the market participants. Traders normally observe the price discovery process by observing through trade book. In some markets, the first few orders of the order book can be observed. Stock exchanges do not disclose the complete order book due to its price sensitivity. Hence, investors cannot fully understand the mechanics behind the price discovery process due to censoring of data by stock exchanges.

In order to make informed decisions, it is important to have information content of both trade book and order book data. This is similar to the role of latent demand in consumer industry. For accurate demand forecasting and its role on pricing of

commodities, retailers should not only research their historical transactions (similar to trade book) but also the latent demand that is captured through consumer surveys (intention to buy is similar to orders in order book that did not result into trades).

Retailers use both sets of information (similar to trade book and order book) for making strategic decisions for optimizing their objective functions. Likewise, in the context of the stock market, we need both order book and trade book data so the market participants can predict the prices accurately; there by helping the markets remain efficient. Deriving a complete order book from visible trades to understand the mechanics behind price discovery process serves as a motivation for the study. Hence, we develop a framework for estimating the order book data to enrich the trade book for better OIB estimation.

1.4 Objectives of the Thesis

Following the motivation derived from the above discussion, the objectives of this study are identified as below:

(i) *This study aims to provide more understanding of the trade direction-based determinants of stock market efficiency to the finance literature. This will be achieved through understanding the impact of order book data on the price discovery process due to the trading activity from both information driven and liquidity driven traders. Such understanding helps us to explain the drivers of stock market inefficiency.*

(ii) *To develop a theoretical estimation method for improving the trade book-based order imbalance.*

(iii) *To develop a model for understanding the predictive power of order imbalance with both complete and missing data of the order imbalance values. .*

(iv) *To measure the predictive ability of the proposed algorithm and provide a comparative analysis of the order book and trade book data based future returns prediction.*

(v) *To help the regulators assess whether market transparency, by disclosing price sensitive order book information, aids or hinders market efficiency.*

In order to achieve the above mentioned objectives, we try to answer the following fundamental research question. *"How can the information content of un-observed or complete order book based orders, be derived for accurate prediction of price movement? "*

One of the main requirements in answering the above question is to source complete order book information. Many machine learning algorithms are used for return prediction in financial time series data. Kokic (2002) uses Multi-layer perceptron (MLP) model for imputing missing values, considering its simplicity, although it is less efficient than popularly used Expectations Maximization Algorithm (EM). Current literature does not provide any basis for using EM algorithm in retrieving hidden orders in stock market data. Given that we cannot source the complete order book and also the fact that the literature does not provide a mechanism to map hidden orders, the central objective is to seek answers for retrieving hidden information relating to orders placed in stock markets. We address this major question by decomposing it into manageable smaller sub-questions as follows:

i. *What features, relating to orders placed by traders, help in predicting future prices? Is it possible to extract the complete order book information from trade book?*

This is the primary question of interest. Given that trade book has lots of missing information with respect to the order book, we need to see what features of trade book will help us to infer the hidden information (that exists in the order book). Existing studies in machine learning literature do not address the level of missing data problem. They use machine learning methods in finance mainly as a comparison tool to standard statistical predictive models (Ahmed et al., 2010).

ii. *What is an appropriate machine learning method for estimating hidden orders based on visible trade book orders?*

In order to retrieve hidden information, we attempt to develop an appropriate methodology that can provide a better estimate for order imbalance (OIB). In the literature, this is normally achieved through some systematic heuristics based pattern recognition approach similar to machine learning methods. This question aims at seeking answers for a) identifying machine learning approach b) developing a comprehensive methodological framework that incorporates all possible factors to make methodology more context-specific to the stock market environment. Existing literature mainly uses neural networks which are hard to interpret in terms of cause and effect relationships (Ahmed et al., 2010).

iii. *How informative is hidden orders data in predicting future price movements?*

This question seeks to answer the economic significance of the methodology. The ultimate objective is to predict short term price movements in stock markets. As prediction of stock prices leads to significant value creation, understanding

the hidden orders (the unmatched, unobserved orders in the trade book) is important part of the research. The information of hidden orders leads one to understand the relationship between these orders, future stock prices and their impact on trade direction.

1.5 Contributions

Order imbalance estimate for price prediction is currently calculated by using trade book data (Chordia et al., 2005). As discussed earlier, trade book data not only suffers from potential misclassification problem but also give incomplete information based estimates (potential under estimation problem). The thesis contributes both to finance and information technology areas by addressing the missing data problem that is apparent for traders using trade book data.

First, for finance research, existing studies based on the role of order imbalance on price movement use incomplete order book that contains only trade book data (Chordia et al., 2002, 2005 and 2009). We innovate by estimating the missing order data and re-examining the role of order imbalance with estimated complete order book. We show that, having estimation of the complete order book can help to predict price direction significantly better than having trade book data.

Second, for the first time, we provide comparative analysis of the future returns prediction results between trade book, our new methodology based estimated order book, and actual complete order book. We show that the results based on the estimated order book are closer to the complete order book. This is a significant contribution to the empirical market microstructure literature in finance. Current empirical studies in market microstructure literature do not address missing data problem. Third, this is the first study on using both estimated and complete order book data. Without a comparative analysis of the complete and estimated data it is

hard to understand the incremental contribution of machine learning methods in advancing finance literature.

For IT literature, first, we provide a machine learning theoretical framework for understanding missing data problem in financial markets. This framework provides a testable statistical procedure for financial securities. Second, this is the first study to utilize Expectations Maximization Algorithm for predicting returns on an intra-day basis to the best of our knowledge. Third, in the current studies on missing data problem, it is not possible to verify the accuracy of the algorithm as the estimates are not directly compared with the hidden data. However, in this case, we have both complete and missing data sets. Hence, it is easy to verify the accuracy of the algorithm based estimated results.

1.6 Organisation of the Thesis

The thesis is organized into six chapters. In Chapter 1, we introduce the terms market efficiency and its inefficiency. We explain the determinants of market mechanism in stock markets. We then give some background on the importance of order imbalance. We compare volume with order imbalance with an example to show why order imbalance is important for price prediction. We use this as the motivation and explain how regulations in stock exchanges censor order book data. We discuss the implications of not having this data available for prediction and how machine learning can help to overcome this censored order book information. The introduction and motivation in this chapter is followed by literature review in Chapter 2.

Chapter 2 presents background literature on the role of information technology in financial markets. We explain order imbalance, order matching

process, price discovery process and missing data problem. We discuss several papers that highlight the major machine learning methods that researchers have used in finance literature. We discuss the shortfalls of the current models used. We provide a comparative analysis to substantiate the usage of Gaussian Process based framework to address our research questions. Finally, we end the chapter with identification of research gaps from the literature review.

In Chapter 3, we provide the theoretical framework for identifying and solving the missing data problem. This is the chapter that integrates finance with IT research. We provide a detailed representation of the missing data problem in finance for development of possible machine learning methodology. We introduce the Relational Markov Network as the procedure to map all orders and their corresponding interactions.

Chapter 4 focuses on developing an empirical implementation strategy for the proposed theoretical idea in Chapter three. It introduces the learning process for mapping missing data by using Expectations Maximization Algorithm. This chapter also provides conditional probability functions for estimating complete order book data.

Chapter 5 presents the evaluations of the proposed method. We provide detailed objective evaluation strategy by focusing on accuracy, efficiency and adaptability as the major dimensions. We clearly map the research questions to the procedure and the outcome. In summary, the outcomes from the learning process can be considered as the hypothesis. For instance, comparison of the estimated order book based results with the complete order book is the outcome in terms of accuracy dimension. Likewise, the duration of the predictability of the estimated trade book data is the outcome of the efficiency dimension. In terms of adaptability dimension,

we analyze the ability of this method to predict several variations across stocks and time horizons.

We conduct several experiments for testing the evaluation based hypotheses. The Australian stock market data are used for conducting the experiments. The first experiment deals with a single stock. This is done for more tractability. The most liquid or heavily traded Australian stock, BHP Ltd., is used to reduce potential estimation errors. In the second set of experiments, several stocks are used by grouping them based on firm size. Overall, the results indicate that, the estimation based trade book data, performs closer to the order book data in the dimensions described in the evaluation strategy.

In Chapter 6, the thesis concludes with discussion of the results and their implications in terms of understanding what EM algorithm requires for prediction to be good, what parameters affect the prediction accuracy. We also discuss the possible direction for future research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

Machine Learning applications in financial services industry are quite diverse. However, most of them are related to developing predictive models. The most commonly covered areas are bankruptcy prediction for firms and financial institutions,[7] predicting credit risk for lending institutions or credit scoring methods,[8] investment portfolio optimization,[9] prediction of foreign exchange rates[10] and stock market movements' prediction.[11] In a survey paper on application of neural networks in finance, Fadlalla and Lin (2001) noted that application of machine learning techniques in finance were virtually non-existent until the early nineties. Initial researchers applied machine learning methods to know whether they are able to predict financial time series better

[7] For example, Kiviluoto (1998); Olmeda and Fernandez (1997); Fletcher and Goss (1993); Leshno and Spector (1996); Tam and Kiang (1992), and Jo, Han, and Lee (1997).

[8] For example, Desai, Crook, and Overstreet (1996); Glorfeld and Hardgrave (1996); Jagielska and Jaworski (1996); Leigh (1995), Piramuthu, Shaw, and Gentry (1994) and Torsun (1996).

[9] For example, Hung, Liang, and Liu (1996), Yeo, Smith, Willis, and Brooks (2002) and Chapados and Bengio (2001).

[10] For example, Yao & Tan (2000), and Kamruzzaman and Sarker (2003).

[11] For example, Chiang, Urban, and Baldridge (1996), Kim and Chun (1998), Teixeira and Rodrigues (1997) on stock market index prediction; Barr and Mani (1994) and Yoon, Guimaraes, and Swales (1994) on stock performance/selection prediction; Wittkemper and Steiner (1997) on stock market risk prediction; Donaldon and Kamstra (1996), and Refenes and Holt (2001).

than the widely used linear models like regression and discriminant analysis (Fadlalla and Lin, 2001).

Information adjustment process in stock markets can be very quick (in seconds or even micro-seconds) and traders who participate in stock market trading are in large numbers (millions for each stock in a day). Therefore, it is computationally challenging to observe the mechanics of price adjustment process. Computational ability can only address the limitations relating to handling voluminous data and understanding the complex patterns in the data series. It cannot address data source related issues such as the quality (completeness) of input data. Most of the stock market prediction is based on trade book data which only has matched order information (trade book does not contain all the orders) which is incomplete.

Conventional methods of conducting research have limitations as addressing this issue of completeness of input data needs a micro approach to understand how markets function during the information adjustment process. The limitations in terms of data availability at transactions level and the availability of computing infrastructure to analyse such order level data can be challenging to ensure market efficiency. This is due to the fact that the traders cannot exploit intra-day, temporal demand and supply-based, stock investor related shocks. If traders could exploit these shocks, there could be faster information adjustment, leading to market efficiency. Understanding the data in their completeness (all orders), needs more cross-disciplinary research between finance and information technology. This has led researchers to embark into cross-disciplinary research that can address these limitations by focusing on developing better predictive

models for standard, everyday stock market data. The usage of machine learning methods in predicting stock market movements is one such avenue.

One of the main objectives of the thesis is to extend this stream of literature by not only using the advantages of higher computational ability to handle voluminous data but also to explore whether data quality matters in better predictions. For instance, missing data problem is a major concern for accurate prediction in financial markets. Trade book contains matched order data which is used by the researchers for prediction. This does not contain information about all buy and sell orders. As discussed in Section 1.2 of Chapter 1, order book data available for traders are very limited. Hence, we undertake cross-disciplinary research by applying machine learning methods to the trade book data in order to extract the features of the order book data. By doing so, we improve the quality of input data so the accuracy of the prediction improves. Understanding the appropriate machine learning method is important for our research.

The objective of this chapter is to review the literature and discuss various machine learning methods for analysing financial time series data. This provides a platform to evaluate various methods and identify the appropriate method to address our research questions. In order to facilitate better alignment of the literature with the research problem, a detailed discussion on the order matching process and the corresponding missing data problem that is being addressed in this thesis is presented. These discussions cumulate into identification of shortfalls in the current methods by highlighting several significant research gaps. These gaps are addressed in this thesis in chapters 3 to 5.

2.2 Speed of Trade Execution in Financial Markets

It is important to understand the evolution of market efficiency in finance literature before we discuss several applications of machine learning methods in finance. Market efficiency is determined by the speed at which information gets adjusted into stock process (Fama, 1970). Over the last five decades, researchers have been attempting to predict stock market movements. However, due to the randomness of information arrival, it has proved to be quite challenging to predict prices that are detected by the random arrival of the information.

Fama (1970) proposed that information arrives randomly and the process of millions of traders reacting and acting based on the random information events ensures that price adjustment to information arrival is almost instantaneous. Hence, it is hard to predict future price movements. This is popularly termed as *Efficient Market Hypothesis* in finance literature (Fadlalla and Lin, 2001). In others words, this hypothesis says that it is in vain to engage in arbitrage strategies or developing predictive models, as in the long run, stock market investment is a zero sum game. Hence, it is hard to consistently create profitable trading strategies based on predictive models. Due to its significant practical relevance, researchers over the last five decades have been testing the relevance of this hypothesis for investment decisions.

The results to test this hypothesis are inconclusive. Several researchers attribute such inconclusive results to limitation of inefficiency (Shleifer and Vishny, 1997) that arise *either* due to market structure that prohibits market to be information wise efficient *or* human behaviour that can influence information adjustment process.

For instance, if there is a price band at which stock has to be traded every day, which is regulated by the stock exchange, then information cannot get fully adjusted as the prices cannot move beyond this band (Deb and Marisetty, 2010).[12] Also, if traders are overconfident on their investment decisions, they may decide not to trade even if the prices need to be adjusted due to the arrival of new information (Daniel et al., 1998).

This debate is still ongoing and quite relevant to this thesis as the thesis aims to understand whether past information related to order imbalance can predict future returns. It is important to note that our research focuses on price adjustment process at very small intervals (seconds and minutes). In such smaller intervals, the chances of identifying inefficiencies in the adjustment process will be much higher due to the physical limits of information adjusted into the prices. For instance, for information to get adjusted into prices investors have to place order with their stock brokers and the stock brokers in turn have to place the orders in stock exchange server. Further, the order should be matched for it to form a trade. Only then, the information gets physically adjusted into the prices. Hence, the smaller the interval, the higher is the likelihood for the information to not get physically adjusted into prices. In other words, the higher is the likelihood to exploit market inefficiency (Chordia et al., 2005). Therefore, we focus our research on these intervals.

2.3 Current Methods for Analysing Financial Time-series Data

Machine learning is a generic term used to refer to several models that use algorithms that allow computers to map and predict outcomes based on some existing empirical

[12] Such rules are popularly called, “price limits”.

data. Alpaydin (2004) includes neural networks, support vector machines, decision trees, and other similar non-parametric probabilistic models under machine learning methods. Ahmed et al. (2010) claim that machine learning methods have established themselves in the last decade as serious contenders to classical statistical models in the area of forecasting. They identify eight main generic machine learning methods, namely, Multilayer Perceptron (MP), Bayesian Neural Networks (BNN), Radial Basis Functions (RBF), Generalized Regression Neural Networks (GRNN), K-nearest Neighbour Regression (KNR), Classification and Regression Trees (CART), Support Vector Regressions (SVR), and Gaussian Processes (GP). Each of these methods is discussed below.

Multilayer Perceptron (MP) is a feed forward neural network method that maps input data to output through multi-layer nodes with each layer connected to the next layer. The advantage of MP is that it can distinguish data that are not linearly separable. It is a very popular network model used both for classification and regression. The introduction of the approximation property in MP, due to which the number of hidden nodes can be controlled, helped in controlling the over parametrization of the model. Hence, the complexity of the model can be controlled by selecting the number of hidden nodes. Hornik et al. (1989) and Funahashi, (1989) showed that MP is capable of approximating any continuous function to any given accuracy, provided that sufficiently many hidden units are available.

Bayesian Neural Networks (BNN) is based on Bayesian probabilistic formulation. Network parameters are treated as random variables having some priori distribution. Low complexity models that produce smooth fits can be dealt favourably

with this distribution. This probability distribution combined with the probability distribution of the new data yields to the posterior distribution that aids in computing prediction (Neal, 1996).

Radial Basis Function neural network (RBF) is an artificial neural network that uses the radial basis functions as their output functions. RBF does not use raw input data, but rather passes a distance measure from the inputs to the hidden layer. This distance is measured from the center value in the range of the variables (sometimes the mean) to give an input value in terms of a Gaussian function. These distances are transformed into similarities that become the data features in a succeeding regression step. The processing of RBF is iterative (Carling, 1992). The outputs of the nodes are combined linearly to give the final output. The width of the node functions (Gaussian function) controls the smoothness of the fitness function (Haykin, 1994; Bishop, 1995).

Generalised Regression Neural Network (GRNN) is a non-iterative method that provides estimates for continuous variables and converges to (linear or non-linear) regression surface (Specht, 1991). The prediction for a given data point is given by the average of the target outputs of the training data points in the vicinity of the given point (Hardle, 1990). The local average is constructed by weighting the points according to their distance from the given point, using some kernel function. The estimation is just the weighted sum of the observed responses or target outputs (Ahmet et al., 2010). This method does not require an iterative procedure and requires only a few training samples to converge.

K-Nearest Neighbor (KNN) method is non-parametric method used for both classification and regression. In this method, prediction is based on the target output of

K training samples (nearest neighbor) nearest to the query point. Euclidean distance is assumed to be the metric used to calculate the closeness between the data points and is computed as the distance between the query point and all the other points in the training set (S B Imandoust et al., 2013). The prediction is the average of the target output values for the closest K training data points.

Classification and Regression Trees (CART) is a recursive, hierarchical, tree-like splitting of input space (Breiman et al., 1984) that builds trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). Input space is divided into smaller samples by repeatedly splitting. A tree will have decision nodes and leave nodes. For a given query data point, the sequence of steps starts at the root node and terminates at the leaf node. This will determine the path of the tree. Splitting recursively will reduce the mean square error until it reaches an acceptable threshold. This splitting will eliminate ineffective nodes and the model complexity is kept in check.

In Support Vector Regression (SVR) method, the threshold of errors is very important than the errors themselves. It mainly helps with problems where the value cannot go beyond a certain threshold limit. Below this error limit, it gives the parameters some space to control model complexity. The input pattern (for which a prediction is to be made) is mapped into feature space. Then dot products are computed with the images of the training patterns. This corresponds to evaluating kernel functions. Finally the dot products are added up using the weights. This, plus a constant term yields the final prediction output (Smola and Schölkopf, 1998).

A Gaussian Process (GP) is a collection of random variables, any finite number of which has (consistent) joint Gaussian distributions. It is fully specified by its mean function and covariance function and defines distribution over functions. This GP will be used as a prior for Bayesian inference. The prior does not depend on the training dataset, but specifies some properties of the functions. Using the prior, a posterior distribution is computed for prediction of unseen test cases (Rasmussen and Williams, 2006).

The preprocessing methods considered by Ahmed et al. (2010), for the time series data are LAGGED-VAL, DIFF and MOV-AVG. In LAGGED-VAL, the input variables are the lagged values and the value to be predicted is the next value. In time series differencing (DIFF) method, forecasting model is applied on the backward difference of the series. In moving averages (MOV-AVG) method, moving averages are computed for the different-sized windows. Table 2.1 shows the performance of the preprocessing methods for the eight machine learning methods.

Table 2.1 Ranking of Pre-processing Methods

Pre-Processing Method	Lagged-Val	Differencing	Moving Averages
MP	1	7	1
BNN	3	6	3
RBF	8	8	8
GRNN	6	2	6
KNN	5	4	5
CART	7	1	7
SVM	4	5	4
GP	2	3	2

Ranks for Lagged-Val and Moving-Avg show similar ranking for all methods. The ranks for Differencing, on the other hand, are different and show worst performance when compared to the other two methods. For this reason Ahmed et al. (2010), ignore this method. Majority of the machine learning applications in finance are based on MP and BNN methods (Ahmed et al., 2010). They conduct a large scale comparison study between the above mentioned eight major machine learning methods by using 3003 economic and financial time series data.

Table 2.2 shows the methods discussed by Ahmed et al. (2010), their key parameters, performance rank order, computational time rank order and overall rank order. Rank order 1 means the highest rank and 8 means the least rank. The key parameters are the ones that control the complexity of the model. Among the eight methods, they find that GP and MP are the best predictors of financial time series data. However, it is important to note that MP is a simple neural network method and hence is limited in its scope to draw meaningful cause and effect relationship between the variables. In other words, it is hard to draw any meaningful economic insights just because the results are statistically significant. Also, there is a need for a clean training dataset for better predictions. GP, on the other hand, overcomes the training dataset related issues by modelling the observed responses of the training data points as a multivariate random variable. The multivariate distribution allows for covering many variables that are normally distributed. For instance, (as discussed in detail in Chapter 3) we model order characteristics and also their interaction characteristics. Both orders and order interactions (that are observed in the training dataset), that are generally random and follow normal distribution, jointly determine the price discovery process.

Table 2.2 Overall Ranking of Eight Major Machine Learning Methods

Model Name	Key Parameter	Performance Rank Order	Computational Time Rank Order	Overall Rank Order
MP	Size of Network	1	2	1
BNN	Size of Network	3	1	3
RBF	Width of Radial bases (beta)	7	3	8
GRNN	Width of the kernels, h	6	6	5
KNR	Number of neighbors, K	5	8	6
CART	Splitting	8	7	7
SVR	Width of the kernels, h	4	5	4
GP	Width of the kernels, h	2	4	2

In addition, GP models update through an inductive approach where additional information is gained during each iterative step and thus such method allows to increasing the estimation power of the training dataset. The posterior distribution of a to-be predicted function value can be obtained using the assumed prior distribution by applying probabilistic manipulation using Bayes rules.[13] Despite their adaptability and applicability to financial time series, GP models are not extensively employed in finance literature.

[13] The technical details of a variant of GP (Expectations Maximization Algorithm) are discussed in Section 4.6.

2.4 Comparative Evaluation of Current Models

In this section, we compare the eight machine learning methods as discussed by Ahmed et al. (2010), based on their capability to predict uncertainty, their ability to define cause-effect relationship, their efficiency in arriving at global minima (convergence), their capability to handle complexity in data and their performance in terms of optimal usage of the resources. These parameters have been selected to establish objective criteria and thereby arrive at an optimal methodology for addressing our research questions. In addition to that, the analysis presented in Table 2.3 helps to justify the rationale for using GP framework in the thesis.

Ahmed et al. (2010) has shown that MP and GP methods are the best out of the eight methods they have studied, to predict financial time series data. As shown in Table 2.3, the evaluation parameters suggests that GRNN and GP methods are both close contenders for the requirement of handling missing data problem and to establish a causal relationship. However, several authors argue that given the real world is dominated by uncertainty, the methodology that does not incorporate uncertainty is of limited value (Ahmed et al., 2010). GP provides an explicit uncertainty measure and does not require the lengthy `training' that is normally required in a neural network methodology. While techniques for obtaining uncertainty from a neural network exist, GP is less complex and better in performance. The main advantage of GP models over GRNN models is that it is a probabilistic model as reported by Ahmend et al. (2010).

Hence, GP outweighs GRNN in performance dimension.[14] In addition to that Ahmed et al. (2010), also report that GP outperforms GRNN for financial data. This could be due to GP's probabilistic model framework that naturally fits into financial forecasting problem as financial models assume probability distribution functions for forecasting future values. Hence, GP outweighs other methods for addressing our research problem.

The other major advantage of GP over GRNN, method is that it is better equipped to handle missing data and it can accommodate both linear and non-linear relationships. Also, due to the Bayesian setting, one may include all possible types of functions that may be derived, and assign a prior probability to each of them such that more plausible functions will have higher probabilities, and to use this knowledge in making a weighted decision. This is an important consideration while dealing with missing data problem in financial markets (Liew Chin Yee and Yap Chun Wei, 2012).

Table 2.3 shows the comparative evaluation of the eight methods used by Ahmed et al. (2010), for comparison on the basis of the five parameters (predicting uncertainty, causal inference, convergence issue, handle complexity and performance) that are essential for financial time series data. Most of the neural networks create a model that maps the input to the output using historical data (supervised learning).[15] Training data are labelled and categorized so the model can be used to produce the output. Uncertainty with the input data is hard to deal with in neural networks and decision tree based

[14] Overall performance, symmetric mean absolute percentage error (SMAPE), an accuracy measure for Lagged-Val pre-processing method for GRNN is 0.1041where as for GP it is 0.0947; that it GP ranks higher than GRNN.

SMAPE for Mov-Avg pre-processing method for GRNN is 0.1033 and for GP it is 0.0962; GP ranks higher than GRNN again.

[15] In supervised learning, inputs are pre-defined and training data are labelled; inputs determine outputs.

models as they are supervised learning models. Hence, in the Table 2.3, we see that MP, RBF and KNN are not very good in dealing with uncertainty. All the other models can predict uncertainty better.

Table 2.3 Comparative Evaluation of ML Methods in Financial Time Series Analysis

ML Methods	Predicting Uncertainty	Causal Inference	No Convergence Issue	Handle Complexity	Performance
MP	No	No	Yes	Yes	Low
BNN	Yes	Yes	Yes	No	High
RBF	No	No	No	No	Low
GRNN	Yes	Yes	No	Yes	High
KNN	No	No	No	Yes	High
CART	Yes	Yes	Yes	Yes	High
SVR	Yes	No	No	Yes	High
GP	Yes	Yes	No	Yes	High

Causal inference tries to look at how well the cause-effect relationships are established and explained by the chosen method. One might argue that, in supervised learning, input causes output; however, it is more deterministic. In the case of our research problem where it is required to establish a causal relationship between order imbalance and future returns, the relationship between order imbalance and future returns is assumed to be probabilistic. Hence, input's effect on output is associated with an element of uncertainty. It is found that generally, models that are based on the Bayesian approach, explain cause-effect relationship between its input and output

variables. The models that are neural network-based are not very good in establishing the relationship.

The methods, MP, RBF, KNN and SVM are not good indicators of cause-effect relationship. Three of the eight methods discussed have convergence issues. MP, BNN and CART have issues with getting stuck at the local minima. Remaining five methods arrive at global minima although some of the SVR has convergence issues (Smola and Scholkopf, 1998). Most of the eight ML methods are good at handling complexity except for BNN and RBF. Performance in the Table 2.3 is based on how long a certain method takes to converge. MP and RBF take a long time to converge whereas the other methods are relatively quicker at converging.

2.5 Order Imbalance

Stock exchanges record transactions in two books: order book and trade book. All buy and sell orders that arrive at the stock exchange are recorded in the order book. Matched orders alone are recorded in the trade book. The difference between the buy orders and sell orders at any given point of time is termed as "*order imbalance"*. If the buy orders are more than the sell orders, there will be positive imbalance. In simple terms, more buys mean more demand; when the demand is high, price will go up. If the sell orders are more than buy orders then there will be negative imbalance. More sell orders mean more supply and less demand. Hence the price is likely to go down. Traders take advantage of the price movement due to order imbalance and make arbitrage profits. Following seminal work by Kyle (1985), researchers attribute such arbitrage

opportunities to temporal imbalances in the orders placed by traders (Chordia et al., 2004, 2005) or simply due to order imbalance.

The following tables (Table 2.4, Table 2.5 and Table 2.6) show a simple example to explain how order imbalance is calculated. It is shown in three steps, each step, referring to the appropriate table. Table 2.4 shows 15 records with date, time, price and bid/ask for a company ABC. Let us say, we are calculating OIB for 5 minutes intervals. The three colours, show the three different 5 minutes intervals. Records represented with yellow are from 10:00:00 am to 10:04:59 am, blue are from 10:05:01 am to 10:09:59 am and green from 10:10:00 am to 10:14:59 am. Under the bid/ask column "A" represents ask and "B" represents bid.

Table 2.4 Order Imbalance Calculation

Company Name	Date	Time	Qty	Bid/Ask
ABC	20120215	10:00:09	85985	A
ABC	20120215	10:01:52	10390	A
ABC	20120215	10:03:05	6500	B
ABC	20120215	10:04:26	15000	A
ABC	20120215	10:05:16	6000	B
ABC	20120215	10:06:39	6557	B
ABC	20120215	10:07:09	10390	B
ABC	20120215	10:08:15	20000	A
ABC	20120215	10:09:10	12800	A
ABC	20120215	10:10:26	6850	A
ABC	20120215	10:11:10	1000	A
ABC	20120215	10:12:15	6600	B
ABC	20120215	10:13:10	15000	B
ABC	20120215	10:14:26	6933	A
ABC	20120215	10:14:50	25000	B

These records are then grouped according to the intervals by summing their quantity for each bid and ask; for the given interval. Table 2.5 shows the grouping. There are two records for each interval; one for ask and another for bid. As shown in table 2.6, SumOfQty of bid is subtracted from SumOfQty for ask to get OIBQty for the interval 10:00. OIBQty is buys (bids) minus sells (ask). Similarly, OIBQty for the intervals, 10:05 and 10:10 are also calculated. In Table 2.6, OIBQty for 10:00 is a negative value (meaning, supply is more than demand), suggesting that the price should fall. For the next interval 10:05, the OIBQty value is positive. Positive OIBQty means more bids than asks; meaning, demand is more than supply. For the interval 10:10 there is positive OIB again.

Table 2.5 Aggregate by Interval

Company Name	Date	Time	SumOfQty	Bid/Ask
ABC	20120215	10:00:00	111375	A
ABC	20120215	10:00:00	6500	B
ABC	20120215	10:05:00	32800	A
ABC	20120215	10:05:00	22947	B
ABC	20120215	10:10:00	14783	A
ABC	20120215	10:10:00	46600	B

On a temporal basis, orders and their imbalance can predict future price movements. Volume can also predict the price as higher volume implies more trading interest and greater price movements. However, higher volume can mean both higher selling (price decrease) or higher buying (price increase).

Table 2.6 Order Imbalance

Company Name	Date	Time	OIBQty (Bid-Ask)
ABC	20120215	10:00:00	-104875
ABC	20120215	10:05:00	-9853
ABC	20120215	10:10:00	31817

Hence, volume cannot indicate the trade direction. Order imbalance, on the other hand, has both magnitude and direction, therefore, has a permanent effect on price formation (Chordia et al., 2002). Generally, the aggregate excess demand of orders is termed as Order Imbalance (OIB). OIB represents the difference between the buy and sell orders at a given point of time. Finance literature has found strong evidence that OIB is a robust predictor in short time intervals (Chordia et al., 2005 and Lawrence et al., 2005). This is quite plausible as the demand shocks in stock inventory affect short-term price movements. For instance, a positive OIB can lead to a temporary increase in the stock price. This increase or decrease in the stock price is not due to matched orders alone. They are due to all the orders that enter the system. This can be understood by the following example:

Let us assume, there is a sell order for $12 and the next order is a buy order for $10. Further, assume that there are 5 buy orders sequentially arranged, at this point of time, there seems to be more demand than supply. By looking at the demand, the seller is not going to reduce his price. But one of the buy orders along the way match up to $12 and a new price is formed at $12. If there were a series of sell orders, then the seller would have to reduce his price to say $10. The new price is $10 now. Hence, all the buy

orders/sell orders influence the price; not just the matched orders. This makes the role of all orders important while considering prediction of future price movements.

Under the circumstances where the transparency regulation does not allow traders to see the order book, the Lee and Ready (1991) method is used to classify the trades. This trade classification can be used to calculate the order imbalance. If the current price of the trade is higher than the previous price, it is marked as a buyer-initiated trade with an understanding that price goes up when demand is more than supply. If the current price is lower than the previous price then it is marked as seller-initiated trade with an understanding that price goes down when supply is more than the demand. This categorization gives us the trade direction. The entire buyer-initiated trades and seller-initiated trades are separately added up for a given time interval. The difference of the aggregated *buys* and *sells* gives the OIB for that time interval. If there are more buys than sells, there is a positive order imbalance and if there are more sells than buys then order imbalance is negative. The positivity or negativity of order imbalance helps us to predict the direction of the price: whether it would go up or go down. This in turn helps to predict the stock return which is the objective of my thesis. Hence, it is very important that these two parameters (Order Imbalance and Stock Return) be measured to be able to predict, particularly when order book data are hidden from the traders.

However, with advancements in technology and competition between stock exchanges, exchanges have started adding attractive features for creating more client base. For instance, ASX maintains only one book that contains all orders entered,

amended, deleted or traded. As discussed in the next section 2.6, all orders are not disclosed to traders. There are hidden orders in the form of undisclosed and iceberg orders types. In many stock exchanges, traders are not allowed to see even such partial order book. They are allowed to see just the trade book. For instance, the NASDAQ Stock Exchange, the New York Stock Exchange, the Singapore Stock Exchange, the Tokyo Stock Exchange and the National Stock Exchange of India do not disclose order book to the traders. None of the major stock exchanges in the world disclose complete order book. This implies that, although, the ASX displays order book to traders, traders cannot estimate the aggregate demand for a given stock. One possible issue for predictive accuracy is the completeness of visible order book. The visible order book to the traders might contain several orders that are either undisclosed or partially hidden (in the form of icebergs).

2.5.1 Order Book Interface in the Australian Securities Exchange

The order book interface of the ASX, shown in Figure 2.1, allows traders to enter buy/sell order, price, quantity, order type and other information. Figure 2.2 shows the order history of a given order. However, all the information is not visible to the traders.

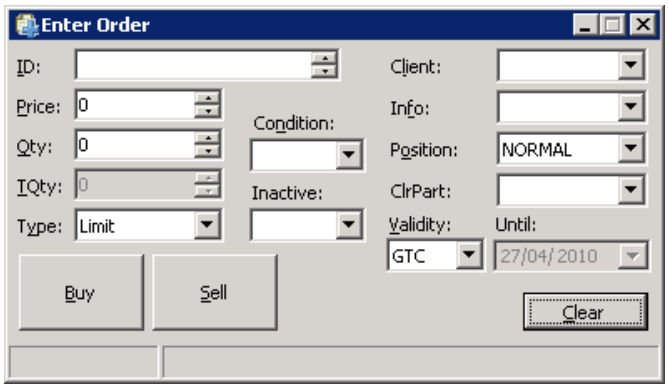


Figure 2.1 Order book Interface

Figure 2.3 illustrates the actual snapshot of the trading interface for Telstra Ltd. The snapshot shows that, the interface displays sell orders (asks in the figure) and buy orders (bids in the figure) of all traders for Telstra stock. The bar chart displays the same information, however, arranged by the order precedence rules of ASX. The bids, represented in green bars, are arranged from the lowest to the highest order price. Likewise, asks, represented in red bars, are arranged in the opposite direction, making sure that the best bid and ask are the highest bid (buy) and lowest ask (sell) prices.

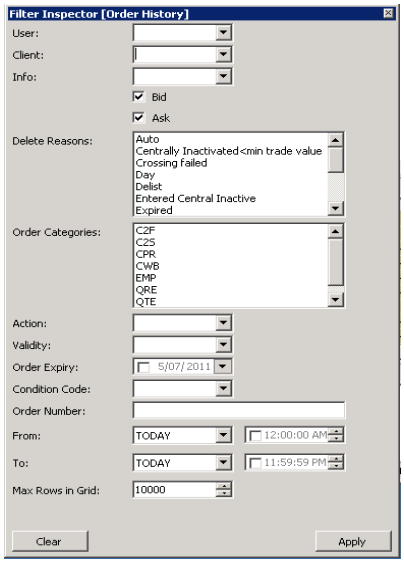


Figure 2.2 Order History Interface

(Source: Australian Securities Exchange)

What is not clear to the trader is that whether there are any other undisclosed orders that are very close to the best bid and ask prices. Such information, although quite valuable, is not visible on the trading interface of ASX.

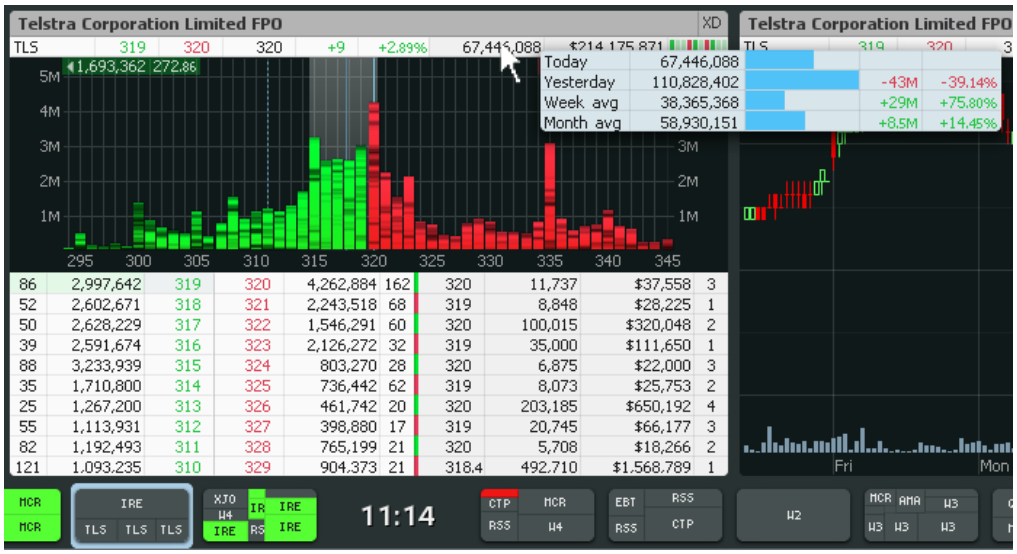


Figure 2.3 Snapshot of the Trading Interface

(Source: Australian Securities Exchange)

2.5.2 Order Book Transparency in the Australian Securities Exchange

As shown in section 2.3.1, the order book stores more information than that is displayed to the traders. More importantly, the Australian Securities Exchange has explicit rules on transparency in the case of two types of order types, namely, Undisclosed Order Type and Iceberg Order Type. We present clear description of these two order types below.

Undisclosed Order Type: These are similar to large block orders. The value of undisclosed order type should be equal to or greater than $500,000. These order type allows the entry of an order where the entire quantity of the order is not disclosed to users. Undisclosed orders are prioritised in price-time priority and may trade against all other orders (including disclosed and undisclosed) in the central order book.

Iceberg Order Type: An Iceberg order allows the trader to disclose only a portion of a given trader's order. Iceberg orders entered by another Participant are not flagged in ASX Trade as iceberg orders. There is no indication in ASX Trade to identify an iceberg

order in the central order book that has been entered by other participants. An iceberg order is entered with a 'shown quantity' and a total quantity.

The 'shown quantity' minimum limit is quantity-based and is set at 5,000 for Equities. There is no limit set on the total order quantity; however, the total order quantity may not exceed 100 times the shown quantity. For example, if the order is entered with a shown quantity of 5,000, the maximum total quantity accepted by ASX Trade will be 100 times the shown quantity or 500,000 in this case.

2.6 Challenges for Prediction with Partial Transparency of Orders

The previous sections point out that all orders are not disclosed to traders. There are hidden orders in the form of undisclosed and iceberg orders types. In many stock exchanges, traders are not allowed to see even such partial order book. They are allowed to see just the trade book. None of the major stock exchanges in the world disclose complete order book. This implies that, although, the ASX displays order book to traders, traders cannot estimate the aggregate demand for a given stock. The extent of disclosure is highly debatable. On one hand, higher disclosure can make markets more speculative. On the other hand, low disclosure increases the cost to traders.

Figure 2.3 that displays the actual screen shot of the visible order book for Telstra Ltd. stock is the actual information visible to traders through a stock broker interface. The X-axis graphs, price data and the Y-axis graph each order size. The green vertical bars represent Bid prices and their red vertical bars represent Ask prices. The best Bid and Best Ask are close to $320 (best bid and best ask appear in the first row of the table below the graph, in Figure 2.3). The aggregate demand or order imbalance can

be calculated by summing all the green bars (bids) and deducting the value with the sum of all the red bars (asks).

The figure indicates that traders can calculate the temporal order imbalance, which can influence the next possible price direction. For instance, higher bids (compared to asks) indicate a possible increase in the future price as there are more people willing to buy compared to sell. One possible issue for predictive accuracy is the completeness of visible order book. The visible order book to the traders might contain several orders that are either undisclosed or partially hidden (in the form of icebergs).

2.7 How Hidden Orders Affect Pricing Strategies?

Figure 2.4 illustrates the role of hidden orders on stock prices. In the figure, the green vertical bars represent bids (buys) and the red bars represent asks (sells). The hidden orders like icebergs are shown below the x-axis. As shown in the figure, $3.6 is the best bid with an order volume of only 500 shares. The next best bid is $3.5, with an order volume of 2000 shares. The best ask is at $3.8 (with an order volume of 1000 shares), however, it is hidden. Likewise, the next best ask, that is hidden is at $4.2 (with an order volume of 500 shares). The only visible best ask is $4.4 with an order volume of 2000 shares (500 visible +1500 invisible shares). The bidder at the best bid, $3.6, can only see the visible ask at $4.4. Therefore, the bidder may need to revise the bid, more aggressively for possible order execution. If the hidden order at $3.8 was visible, the bidder can exercise a less aggressive pricing strategy. Hence, hidden orders can lead to possible mispricing due to the invisibility of potential better bids or asks. Figure 2.4 shows that the bid order at $3.2 is partially hidden. Likewise, the next bid at $3.0 is

completely hidden.

Figure 2.4 Illustration of Order Book with Hidden Orders

Such partial or hidden order book gives rise to potential problems associated with execution of orders. Traders cannot gauge the actual market demand and supply forces for strategically pricing their orders. This issue leads traders to wrongly judge the market movements. Hence, it is hard to predict future returns with the observed partial order imbalance. In summary, hidden orders play a significant role on pricing strategies of traders and the corresponding price formation.

2.8 How Hidden Orders are Incorporated in the Estimation Process?

As shown in the previous section, missing information related to orders, similar to Iceberg orders can mislead traders to predict market movements. Hence, such hidden

data need to be accounted for estimating the aggregate market demand. Following finance literature, we use OIB as the main predictor of short-term price movements. Given that some orders are hidden, estimating the actual OIB is the main challenge that is being addressed in this thesis. The hidden orders need to be estimated for calculated the actual OIB of a given stock. Otherwise, prediction of future price movement will be biased and hence costly for traders.

Consider the Figure 2.4; the actual order book-based OIB is 3250 shares (6750 shares of bid orders minus 3500 shares of ask orders). Traders would estimate the OIB from visible orders as 3750 shares (4250 shares of visible bid orders minus 500 shares of visible ask orders). With higher visible OIB, traders' order placing and pricing strategy would vary. Such wrong estimations can lead to order mispricing and orders not being able to execute successfully. The hidden orders would compete and move the prices beyond the reach of an ordinary trader. Hence, including such hidden orders in the estimation part is critical for more accurate prediction of future price movements. In the following section, we will explore how the orders are matched and how a new price is formed.

2.9 Order Matching and Price Discovery Process

The order book contains all orders placed along with their respective order attributes. The primary attributes of orders are type (buy or sell), price, time and quantity. There could be many secondary attributes that give more information on orders. The main secondary attributes are, trader type (for example, institutional or retail), order execution instruction (for example, is there any price limit for order execution or the order has to

be executed at the current market price) and broker identity. Majority of the stock exchanges follow price, time and quantity as the order precedence rule. This implies that if there are two orders at the same price then the order that has arrived first is given preference in execution and if two orders of same price arrive at the same time then the order that has quantity closer to the order on the other side of book is given preference. Thus order precedence algorithms help in matching buy and sell orders.

Figure 2.5 describes the order matching process that happens in majority of electronic stock exchanges. This figure depicts two boxes that represent order book and trade book. Order book at time "*t*" contains four sell orders (SO1, SO2, SO3, SO4) and two buy orders (BO1 and BO2). As shown in the figure, at time "*t*" only one buy order and one sell order is matched through the order precedence algorithm. For instance, say, there are three attributes: price, trader type and quantity on which the orders are matched. Some of the orders get matched on these attributes and enter into trade book; however, some orders remain unmatched at a given point of time. The remaining orders that are not matched are carried forward to time "*t+1*". The trade (matched orders) at time "*t*" (SO1 and BO2) is the only entry that goes into trade book (T1). At time "*t+1*" order matching can happen either due to the arrival of new orders, due to the revision of the unmatched orders that belong to time "*t*" or incomplete matching of orders, in terms of order size, can lead to further execution of unmatched orders at time *"t+1"* . For instance, if order size of SO1 (say, sell order of 10 shares) is larger than BO1 (say, buy order of 4 shares) then unmet part of the SO1 order (remaining 6 shares that are unmatched after 4 are matched with BO1) are carried into order book at time *"t+1*".

Trade T2 at time "*t+1*" is due to arrival of a new buy order BO3 at time "*t+1*" matching with previously unmatched SO3 that arrived at time "t". However, trade T3 occurs at time "*t+2*" is due to the revision of orders that entered the order book at time "*t*", namely, BO1 and SO5. Some orders wait for the right match and trade T4 occurs when such orders are matched at time *"t+3"*. The price difference between matched orders (trades) at time "*t+1*" and time "*t*" represent the trade direction at time "*t+1*". That is, if the price at "*t+1*" is higher than at "*t*", there is an upward direction in price and vice versa. This process of *changing prices to discover new prices* is often termed as *price discovery process* in stock markets.

The description of order matching process highlights two important issues. First, how traders organize to trade and how their organizational structure influences market price discovery process is mainly captured through the orders they place at a given point in time. Traders' organizational dynamics are observed through orders, their corresponding attributes and resultant order imbalance at a given point of time. As shown above, order imbalance at time "*t*" has an effect on orders and trades at time "*t+1*". Also, it indicates that order imbalance magnitude (how high/low the imbalance is) plays an important role on the price discovery process. For instance, a higher positive order imbalance at time "*t*" implies higher likelihood of price increase at time "*t+1*". Hence, there is a possibility to classify order imbalance levels based on their effect on the prices and corresponding trades. For instance, a high level of positive order imbalance (more buy orders than sell orders) can trigger arrival of new sell orders from arbitrageurs to push prices back to their equilibrium values.

In such instances, order interactions happen mainly due to new buy order arrivals which will interact with all possible sell orders on the book due to excessive demand from the arbitrageurs.

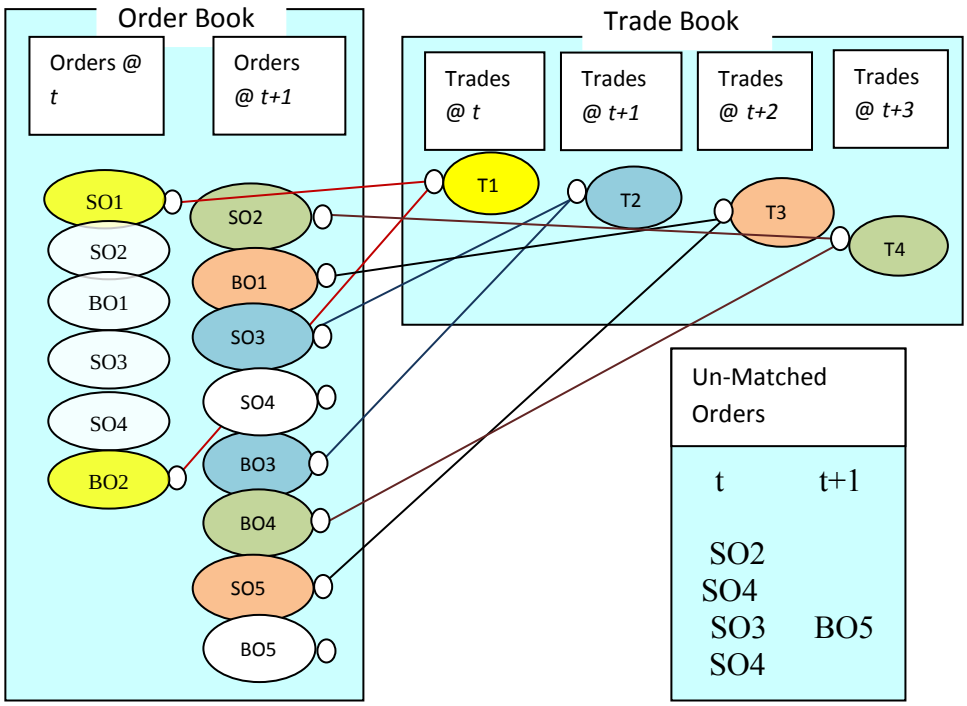


Figure 2.5 Order Matching Process in Electronic Stock Exchange

Hence, order interactions are related to a specific level of order imbalance. Second, the complete organizational structure of traders' interactions can be captured by including both matched and unmatched orders. As indicated in Figure 2.5, trade book depicts partial data and hence it is incomplete. It is important to note that these unmatched orders also indirectly contribute to the price discovery process as they are part of the actual order imbalance at that point of time. However, they are not observed due to stock exchange regulations. Only matched orders are observed by traders. Hence,

the observed data suffers from *missing data problem* and affects predictive ability of orders based information. The missing orders data, as shown in the right-most box of Figure 2.5, needs to be included to draw more meaningful inferences on trade direction. The focus of this thesis is to design and implement a procedure for estimating the unobservable order book data. The objective is to map orders that are hidden in order book and thereby estimate actual order imbalance (OIB, calculated from order book) compared to observed traded order imbalance (OIB calculated from trade book).

2.10 Missing Data Problem

Inferring trade direction by observing only the trade book data attracts missing data problem for predicting trade directions. Missing data are generally classified into three types (Little and Rubin, 1987), namely, Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). In our context, traders (or stock market analysts) are unable to see complete data for inferring trade direction due to the data being censored by the stock exchanges. In such a case, the observed data contains only those orders that dominate other orders (as only best bid and best ask are executed and transferred into the trade book). All other orders that might influence the future trade direction are not visible.

In other words, trade direction or price movement (the dependent variable) can be influenced by latent variables (hidden orders that are not matched and not transferred into trade book) which are missing in the sample (trade book). It might appear on the surface that data are missing not at random as the stock exchange stops certain data from being seen. However, what is visible to the traders (observed data in the trade book) is

due to the random interaction of orders that is beyond stock exchanges' control. Therefore, seemingly non-random censoring gives the illusion of non-random event (MNAR), the fact that the order interactions happen randomly leads us to classify data as Missing At Random (MAR). The data cannot fall under the category of MCAR due to censoring by the stock exchanges. Hence, in the context of stock market, data are missing at random (MAR).

Hence, trading algorithms that are used by investors with trade book data as the input can lead to biased estimates due to missing data problem. This happens due to under-representation of the explanatory variables. Given that, while inferring trade direction, it is quite costly to ignore the missing data, the bias can be reduced by imputing missing values. The imputation methods can range from simple mean imputation to highly advanced machine learning techniques (Gelman and Hill, 2007). The appropriateness of a given method mainly depends on value gains from trading strategy, cost of inaccuracy and implementation viability.

Our research aims to jointly model order attributes like order size and order value[16] of unmatched orders (which we call as missing data) and matched orders for better predictions. We propose to use a probabilistic model for estimating the joint distribution of orders. This is discussed in detail in the framework section of Chapter 3.

[16] We may consider including more attributes including type of investor placing orders, whether the order is market or limit order. However, the final set of order attributes will be decided after analysing the information content and corresponding influence of each attribute on the price discovery process.

2.11 Shortfalls of Current Methods

The existing literature suffers from two important shortcomings. First, none of the machine learning applications in finance addresses training dataset related issues. In other words, they do not check the quality of input data. Hence, the predictions made are not with a good representative training dataset which is often noisy, leading to potential misleading predictions. Stock exchanges censor order book information as this is price-sensitive. Given that, the quality of the input variables is the main determinant of the quality of the predicted output, the problem becomes acute in stock market setting. Second, existing studies use machine learning methods, mainly as a tool and they do not focus on developing a methodology that incorporates several nuances related to financial markets that can characterize specific data generating process related to a given market. For instance, they assume that the statistical distributional properties that are applicable to the machine learning method are similar to input data distribution. There is no serious attempt to characterize and align the distributional properties to a specific machine learning methodology. Ignoring this issue can lead to biased results.

As discussed in section 1.2.1, order imbalance (OIB) plays a very important role in price prediction. OIB calculated from order book can give accurate prediction when compared to OIB calculated from trade book due to trade book's limited information content. As discussed in section 2.9.1, for stock market data, missing data problem is associated with incompleteness of trade book data due to missing unmatched orders in OIB.

2.12 Research Gaps

Application of machine learning methods in financial services related issues is quite broad. Most of the applications are related to developing predictive models. There is limited literature, (next to none) about the research on incompleteness of the input time series data. There is not any published work on exploiting powerful machine learning algorithms on intra-day stock market data. The following gaps have been discovered in the current literature:

i. A formal machine learning methodology that properly defines data generating process, which matches the distributional properties of the input data, which can address missing data problem (to be more specific to missing orders data problem) is not developed for financial time-series related research.

ii. Existing financial markets based studies do not investigate and explore the information content of order book or hidden orders in predicting price movements.

iii. Current financial markets based studies mainly use neural networks for predicting price movements. However, Ahmed et al. (2010) note that neural networks based methods, compared to Gaussian Process (GP) based machine learning methods, are inferior in predicting financial time series. There are no studies that use GP[17] based machine learning approach for stock market predictions.

[17] Appropriateness of Gaussian Process (GP) is discussed in detail in Section 2.2.

iv. Computational ability of existing algorithms can only address limitations related to handling time critical voluminous data and understanding complex patterns. It cannot address input data quality related issues.

v. Existing studies on financial market do not focus on the issues related to handling problems due to missing input data.

The thesis addresses these five important research gaps by a) developing a formal theoretical framework with clear data generating process and distributional properties of orders and trades in stocks markets; b) investigating the role of hidden orders on future prices and whether they have prediction power; c) using Gaussian Process based machine learning methods; d) Given that our methodology incorporates new information related to hidden orders, the quality of input data will be better than the existing studies; and e) Our methodology, as described in Chapter 3, uses Expectations Maximization Algorithm specifically to handle the missing data problem.

2.13 Summary

In this chapter we provided a discussion of the literature in both finance and information technology areas that is relevant to our research questions. We argue and provide justification on why inter-disciplinary research needs to be carried out to understand the role of order imbalance to predict stock returns. We highlight and evaluate various machine learning methods to justify the usage of Gaussian Processes frame work for our research. We identify five major research gaps, namely, lack of clear theoretical underpinnings for the usage of machine learning tools in finance; limitations of using

trader book data in predicting stock returns; no clear justification on the usage of a particular machine learning method in finance; dearth of research in finance to address the input data quality while using machine learning methods and lack of research on how missing data problem can be addressed in finance.

Chapters 3, 4 and 5 provide a clear road map for filling the above mentioned five gaps. In this process of seeking answers to our research questions, we provide theoretical foundations to describe the data generation process and distribution properties of various parameters that are related to the prediction of stock returns in chapter 3. In Chapter 4, we present the learning process to address the missing data problem and to improve stock returns' prediction. Chapter 5 presents experiments and the performance evaluation framework and analysis. Thus, the thesis takes the lead from the research gaps in this chapter and provides a process, chapter-wise, to make an accurate stock returns' prediction based on trade book.

This page is intentionally left blank.

Chapter 3

MISSING DATA MECHANICS IN RETURN PREDICTION

3.1 Introduction

In majority of the stock exchanges, due to the regulations on transparency of trading information, complete order book is not visible to traders. In addition, trading information such as price or trading volume is observed with a time delay. These limitations reduce the predictive power of the return prediction algorithms. This chapter describes how these limitations can be overcome by modeling trade book's missing information and how it can be characterized with data and distributional assumptions.[18] The importance of order book information in predicting future returns is discussed in detail in this chapter.

Estimating from missing data is the first step in estimating the order imbalance. One needs to identify the characteristics of such missing data in order to choose the method for estimation. Hence, the main objective of this chapter is to show how the observed information (of order book) can be leveraged to enhance the price movement predictions from the unobserved data (or incomplete data) of the trade book. Existing literature does not provide a theoretical basis for justifying the required estimation procedure for accurately predicting with missing data. This chapter provides a detailed

[18] Stock market data are assumed to have normal distribution (Campbell, Low and MacKinlay, 1987).

theoretical estimation procedure for missing data along with a running example to demonstrate that estimation accuracy can be improved by applying maximum likelihood procedure, by incorporating missingness as part of the estimation procedure.

The chapter is mainly divided into three parts. In the first part, a discussion on the features of observable and unobservable orders (such as mean and covariance) that are required to predict stock returns is initiated. Further, the discussion moves on to understand how orders placed by traders and the imbalance of buy and sell orders is justified as a predictor of stock return. Observed information in the order book constitutes completeness of information and the unobserved information in the trade book constitutes the missingness. Incomplete information of orders in trade book results in underestimation of order imbalance in the incomplete dataset (order imbalance calculated from trade book). Hence, by borrowing the observed information from order book, the prediction bias is reduced.

In the second part, the chapter introduces missing data classification methods to show that the missingness of the stock market data falls under the missing at random (MAR) classification so the methods used throughout the thesis are consistent with MAR classification.

In the third part, a detailed discussion on the distributional properties and estimation methods that are required to recover the missing data with respect to stock market data are presented. The mechanism for addressing the missing data problem is explained with a running example that shows how the proposed estimation method provides superior estimates and thereby improves the prediction accuracy.

3.2 Major Predictor of Short Term Price Movements

Traditionally, stock exchanges normally maintain two electronic books, namely, order book and trade book. The order book contains information of all the orders entered by traders and trade book contains only the matched or executed orders. If a given sell order (bid) and buy order (ask) match, then such matching, termed as a trade enters into the trade book. Hence, by construction, the trade book contains partial information with only those orders that are the best sell and best buy[19] orders at a given point of time.[20] The matching process increases the chances of movement of prices, as the investors have to discover a new price based on the leftover orders or arrival of new orders into the order book. Hence, orders that are not visible, due to non-availability in the trade book, have significant influence on price changes or returns. As investors' main objective is to determine future returns, orders information becomes crucial in prediction.

Generally, the aggregate excess demand of orders is termed as Order Imbalance (OIB). As discussed in detail in Section 2.5, OIB represents the difference between the buy and sell orders at a given point of time. Finance literature has found strong evidence that OIB is a robust predictor for returns in short time intervals (Chordia et al., 2005). This is quite plausible as the demand shocks in stock inventory affect short-term price movements. For instance, a positive OIB can lead to a temporary increase in the stock price. The positivity or negativity of order imbalance helps us to predict the direction of

[19] Best buy and best sell orders are the orders that have been matched or executed based on the priority in the order book queue.

[20] Even in stock exchanges that do not maintain two separate books, namely, order book and trade book, complete information is not visible to traders due to several types of hidden orders that are discussed in Chapter 2.

the price: whether it would go up or go down. This in turn helps to predict the stock return which is the objective of the thesis. Hence, it is very important to measure order imbalance in order to predict stock returns accurately. However, the estimates of order imbalance potentially suffer from underestimation problem if complete order book is not visible to traders as the cardinality of trade book and order book is different. Therefore, the process of order imbalance estimation, to reduce the underestimation problem can be classified as *missing data problem*.

Dealing with under estimation problem due to missing data with respect to stock market prediction (or forecasting) generally requires the distributional properties of the data. For example, in case of normal distribution, we need mean, variance and covariance (denoted by μ, σ^2, Σ respectively) of the sample to estimate the log likelihood value that produces the best fit to the data. These parameters need to be estimated from the data. According to the financial economics literature (Campbell, Low and MacKinlay, 1997), returns distribution is assumed to have normal distribution. In complete datasets (as in the case of OIB calculated from the observed orders of the order book), these parameters are assumed to be known. In datasets with missing data (OIB calculated from trade book), these parameters are not known. In such instances, arriving at unbiased estimates is a challenge for researchers.

3.3 Types of Missing Data

Unplanned missing data are potentially damaging to the validity of a statistical analysis. Rubin's (1976) theory describes situations where missing data are relatively benign. Missing data are generally classified into three types (Little and Rubin, 1987), namely,

Missing At Random (MAR), *Missing Completely At Random* (MCAR) and *Missing Not At Random* (MNAR). In the context of this thesis, data are missing mainly by censoring by the stock exchanges. In such a case, the observed data contain only those orders that dominate other orders (as only best bid and best ask are executed and transferred into the trade book). All other orders that might influence the future trade direction are not visible. In other words, trade direction or price movement (the dependent variable) can be influenced by latent variables (hidden orders that are not matched and are not transferred into trade book) which are missing in the trade book.

Dataset to explain the data classifications

The dataset provided in the Table 3.1 are a representative sample of a typical order book and trade book imbalances, along with corresponding stock returns. This data are used to explain different data classifications (MAR, MCAR and MNAR) and further used as a running example throughout this chapter in building the predictive model parameters.

Table 3.1 shows a representation for the stock AGK (AGL Energy Ltd.) on 4th Jan 2012. First column represents complete order book imbalance (OIB), second column represents the return calculated as a dollar change in price and the third column represents the order imbalance calculated from trade book. Section 2.5, in the previous chapter, explains in detail how OIB is calculated from order book and trade book. Let us take the instance of the first row where the complete order book imbalance (OIB) is 807 orders. This value is the difference between buy orders and sell orders in order book. The corresponding return for the first row OIB data is 0. This implies that there is no change in the price compared to the previous price. The corresponding trade book order

imbalance (TOIB) is 272 orders. This value indicates the difference between the buy orders and the sell orders in the trade book. It is evident that this value 272 of TOIB is less compared to 807 of OIB. This disparity is due to the fact that the trade book records only matched orders. Hence, the information available in trade book is partial as that of matched orders alone.

Table 3.1 An Example of Orders Dataset

OIB	Return	TOIB
807	0	272
692	0.59543	346
884	0.590361	380
385	0.574309	271
623	0.582243	162
635	0.58292	229
461	0.581779	201
340	0.581565	136
464	0.5821	236
337	0.582637	176
519	0.583175	243

Using the above representative data from Table 3.1, the next subsections explain how missing data can be classified in machine learning literature.

3.3.1 Missing at Random (MAR)

Data are missing at random (MAR) when the possibility of missing data on a variable *Y* is related to some other measured variable (or variables) in the analysis model but not to the values of *Y* itself. The term missing at random (MAR) is somewhat misleading because it implies that the data are missing in a haphazard fashion. However, MAR actually means that a systematic relationship exists between one or more measured variables and the probability of missing data.

For example, consider the dataset provided in Table 3.1. OIB is calculated from complete data in the order book and hence considered complete. However, TOIB is calculated from matched orders alone and hence it is considered to be incomplete data. The missingness in TOIB is not due to the orders in trade book. In other words, missingness in TOIB is not due to TOIB itself; it is due to properties of the order that are triggered due to stock exchange matching conditions, i.e., missingness in TOIB depends on orders that are not qualified to enter in the trade book due to censoring by stock exchanges. Missingness of the data is not due to orders themselves but due to another variable (stock exchange rules) that is unrelated to the orders. Please note that MAR does not imply that the data is completely random. It only means, missing data is due to observations that are not related to the data. Hence, the data used in the thesis can be considered as missing at random (MAR). This explanation satisfies "*Rubin's (1976) definition of MAR*".

3.3.2 Missing Completely at Random (MCAR)

MCAR assumes that missingness is completely unrelated to the data. If data is missing at random with no pattern (not related to any related or unrelated observations) then it is classified as missing completely at random (MCAR). However, if the data is missing randomly due to unrelated observations then it is missing at random (MAR). Please refer to Luck and Rubin (2002), Allison and Graham (2002) and Sclomer, Bauman and Card (2010).

As an example, refer to the Table 3.2, where the first column has OIB data (complete data); third column MTOIB has some of the values of TOIB (MAR data from column

4), randomly deleted. As per the definition of MCAR, the observed data are a random sample of hypothetically complete data in fourth column (TOIB). This column represents MCAR for the table data.

One possible test for identifying whether the missing data should be classified as MCAR is a mean difference test. To test if the data are MCAR, complete data needs to be separated from missing data. In this case, OIB from order book and trade book (TOIB) need to be separated and their group means are calculated and compared. If the difference is small, it suggests that the two groups are randomly equivalent and therefore the data are MCAR. If the group mean difference is large then the two groups are systematically different.

Table 3.2 An Example of MCAR and MAR Data

OIB	Return	MTOIB (MCAR)	TOIB (MAR)
807	0	272	272
692	0.59543	346	346
884	0.590361	----	380
385	0.574309	271	271
623	0.582243	----	162
635	0.58292	229	229
461	0.581779	201	201
340	0.581565	136	136
464	0.5821	----	236
337	0.582637	176	176
519	0.583175	243	243

3.3.3 Missing Not at Random (MNAR)

When missing values on a variable *Y* are related to the values of *Y* itself, then it is classified as missing not at random (MNAR). For instance, if the missingness in TOIB is due to TOIB itself, then it is MNAR. Unlike MAR mechanism, there is no way to verify whether the data are MNAR without knowing the values of missing data.

3.4 Missing Data Mechanism

In this thesis, as discussed in the Section 3.2.1, due to its appropriateness, the data are assumed to be MAR. MAR mechanism occurs when the probability of missing data on a variable *Y* is related to another measured variable in the analysis model but not to values of *Y* itself.

This can be represented as

$$p(R/Y_{obs}, \emptyset) \tag{3.1}$$

where p is the probability distribution, R is the missing data indicator, that indicates whether the data are missing or not, Y_{obs} is the observed parts of the data and $\emptyset$ is a parameter (or set of parameters) that describes the relationship between R and the data. Equation 3.1 means that the probability of missingness depends on the observed portion of data via some parameter $\emptyset$ that relates Y_{obs} to R. In other words, probability of missing data on Y can depend on Y_{obs} and not on Y missing. The figure 3.1 below shows MAR mechanism.

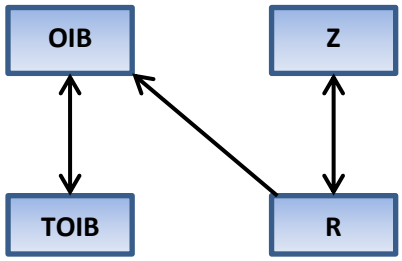


Figure 3.1 MAR Mechanism

In the Figure 3.1, there is an arrow pointer to OIB from *R* but no arrow from *R* to TOIB. This means that the probability of missing orders depend on the complete order book that contains all orders (OIB) and does not depend on only the matched orders (TOIB). Therefore, *R* depends on *OIB* but not *TOIB. Z* represents a spurious relationship that occurs when *R* and OIB are mutually correlated with one of the unmeasured variables in *Z*. This representation satisfies Rubin's (1976) definition of MAR.

3.5 Methods to Address Missing Data Problem

Researchers have proposed various techniques to address missing data problem. Traditionally, the two popular methods of dealing with missing data are, removing the missing cases and filling the missing values.

Two of the most common deletion methods are list-wise deletion and pairwise deletion (Peugh & Enders 2004). List-wise deletion discards the data for any case that has one or more missing values. Pairwise deletion attempts to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis. The primary advantage of these methods is that they are convenient to implement and are standard options in statistical

software packages. However, these approaches assume MCAR data and can produce distorted parameter estimates when this assumption does not hold. Even if the MCAR assumption is plausible, eliminating data is wasteful and can dramatically reduce power of prediction. Consequently, there is little to recommend these techniques unless the proportion of missing data is trivially small (Wilkinson & Task Force 1999).

Filling in the values is also called imputing. Imputation is an attractive strategy because it yields a nearly complete dataset. At first glance, imputation is also advantageous because it makes use of data that deletion approaches would otherwise discard. Despite this advantage, most of the approaches produce biased parameter estimates even when the data are MCAR. Stochastic regression imputation, a single imputation technique, is the sole exception. Stochastic regression method gives unbiased parameter estimates for MAR data (Little & Rubin 2002). However, analyzing single imputed dataset effectively treats the filled-in values as real data, so even the best single imputation (e.g., stochastic regression imputation) will underestimate the sampling error. Although this method is good, it has problems that make it inferior to maximum likelihood estimation and multiple imputation (MI) methods. Multiple imputation (Rubin 1977, 1978) creates several copies of the dataset and imputes each with different plausible estimates of the missing values. Unlike the single imputation method, MI is a method that overcomes the underestimation issues by appropriately adjusting the standard errors of the missing data. Hence, for the analysis in the thesis, both maximum likelihood estimation and multiple imputation method are used. In the next subsection, we learn more about maximum likelihood estimation process in calculating the

maximum likelihood estimate. Multiple imputation method is discussed in detail in the methodology section in Chapter 4.

3.5.1 Maximum Likelihood Estimation (MLE)

Maximum likelihood estimation plays a vital role in missing data analyses by producing best fit to data (Schafer & Graham, 2002). The starting point of the maximum likelihood analysis is to specify a distribution for the population data. The general assumption by researchers is that the data are normally distributed. Although the normal distribution plays an integral role throughout the estimation process, the basic mechanics of estimation are largely the same with other population distributions. The mathematical approach of maximum likelihood relies heavily on probability density function that describes the distribution of population data. In other words, the density function describes the relative probability of obtaining a score value from normally distributed population with a particular mean and variance.

The multivariate normal distribution generalizes the normal curve to multiple variables. The probability density function for the multivariate normal distribution is:

$$L_i = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{-.5(Y_i-\mu)^T \Sigma^{-1} (Y_i-\mu)} \quad (3.2)$$

Y_i, the score vector has k number of individual scores,[21] μ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix. The key portion of the formula is $(Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)$, is the Mahalnobis distance (Mahalnobis, 1936). This Mahalnobis distance term quantifies the standardized distance between an individual's score point and the center of the

[21] For example, values for a given OIB and TOIB.

multivariate normal distribution (mean). The collection of the terms left of the exponent term in equation 3.2 is a scaling factor that makes the area under the distribution sum (i.e., integrate) to 1. The highest point on the distribution (peak) corresponds to the score that is exactly equal to the population mean. Small variations between score vector and the mean vector produce large likelihood values, whereas large deviations yield small likelihoods. *This implies high likelihood of the estimates to the population mean can be obtained when the observed mean is close to the population mean*. This will be cross referenced when the analysis is complete to explain clearly if the method yields to the expectation set here.

3.5.1.1 Computing Individual Likelihoods

The multivariate normal density describes the relative probability of drawing a set of scores from a multivariate normal distribution with a particular mean vector and covariance matrix. To illustrate the computations, consider the OIB and TOIB scores from Table 3.1. For the sake of demonstration, assume that the population parameter values are as follows:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_{\boldsymbol{OIB}} \\ \mu_{\boldsymbol{TOIB}} \end{bmatrix} = \begin{bmatrix} 513 \\ 268 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\sigma^2_{OIB}} & \boldsymbol{\sigma_{OIB,TOIB}} \\ \boldsymbol{\sigma_{TOIB,OIB}} & \boldsymbol{\sigma^2_{TOIB}} \end{bmatrix} = \begin{bmatrix} 32277.69 & 1850.93 \\ 1850.93 & 15222.36 \end{bmatrix}$$

To begin, consider the orders; whose individual OIB is 884 and TOIB rating of 380. Substituting these scores in equation 3.2 yields a likelihood value of -14.31 as follows:

$$L_i = \frac{1}{(2\pi)^{\frac{2}{2}}\left|\begin{matrix}32277.69 & 1850.93\\ 1850.93 & 15222.36\end{matrix}\right|^{\frac{1}{2}}} e^{-.5\left[\begin{matrix}884\\380\end{matrix}\right]-\left[\begin{matrix}513\\268\end{matrix}\right])^T\left[\begin{matrix}32277.69 & 1850.93\\ 1850.93 & 15222.36\end{matrix}\right]^{-1}(\left[\begin{matrix}884\\380\end{matrix}\right]-\left[\begin{matrix}513\\268\end{matrix}\right])} = \text{-14.31}$$

Visually, the likelihood is the height of the normal distribution at a point where scores of 884 and -14.31 interact. After computing the likelihood estimates, the natural logarithm of individual likelihood estimates[22] is calculated to simplify the mathematics of maximum likelihood.

3.5.1.2 Computing Log Likelihood

The individual log-likelihood estimates for multivariate normal data (Sprott, 2000) can be represented as

$$log\ L_i = \log\{\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} e^{-.5(Y_i-\mu)^T\Sigma^{-1}(Y_i-\mu)}\} \tag{3.3}$$

Where the terms in the braces produce the likelihood value for case *i*. After distributing the logarithm, the individual log-likelihood becomes

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y_i-\mu)^T\Sigma^{-1}(Y_i-\mu) \tag{3.4}$$

The log likelihoods have the same meaning as individual likelihoods. The estimation routine for estimating the parameters (mean vector and covariance matrix) of the multivariate normal distribution repeats the log-likelihood computations many times, each time with different random estimates[23] of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Each unique combination of

[22] The individual likelihoods are very small numbers and hence are not tractable. Converting into natural logs ensure that they are more tractable for further calculations.

[23] Some of the model fitting programs tend to use regression procedure while other optimization routines use calculus derivatives to adjust the parameters and improve the log-likelihood to facilitate the estimation process.

parameter estimates yields a different log-likelihood value, and the goal of estimation is to identify the particular constellation of estimates that produce the highest log likelihood and thus the best fit to the data.

Maximum likelihood estimation is actually far more flexible because the mean vector and covariance matrix can be functions of other model parameters. For example, a multiple regression analysis expresses the mean vector and the covariance matrix as a function of the regression coefficients and a residual variance estimate. However, estimating complex models involves a collection of equations each of which contain a set of unknown parameter values. Therefore, complex applications of maximum likelihood estimation generally require iterative optimization algorithms which will be discussed in detail in Chapter 4.

3.5.2 Maximum Likelihood for Missing Data

As discussed in the Section 3.2.1, the maximum likelihood estimation repeatedly auditions different combinations of population parameter values until it identifies the particular constellation of values that produce highest log-likelihood value. Conceptually, estimation process is the same with or without missing data. However, missing data introduce some additional nuances that are not relevant for complete-data analyses.

Incomplete data records require a slight alteration to the individual log-likelihood computations to accommodate the fact that individuals no longer have the same number of observed data points. Missing data analyses require iterative optimization algorithms, even for very simple estimation problems. Methodologists, currently, regard maximum

likelihood as a state-of-art missing data technique (Schafer & Graham, 2002) because it yields unbiased estimates under a missing at random (MAR) mechanism. From a practical standpoint, this means that maximum likelihood will produce accurate parameter estimates in situations where traditional approaches fail. Even when the data are missing completely at random (MCAR), maximum likelihood will still be superior to traditional techniques (e.g., deletion methods) because it maximizes the statistical power by borrowing information from the observed data. This can be further illustrated by considering the distributional properties of the data and comparing it with distributions after random deletion.

3.5.2.1 Missing Data and Distributional Properties

Understanding the distributional properties and their underlying assumptions can help in estimating the relationship between OIB and future price movements. However, while implementing this idea, researchers have to deal with missing data and a procedure to impute the missing values with certain assumptions relating to distribution properties. There is a chance that the distributional properties of missing and imputed data may vary. Also, the extent of missing data may affect the distribution properties. For instance, 10% of missing data may have a different distribution compared to the distribution of 20% missing data. Likewise, a stock that is large and has many trades would have more continuous observations compared to a small stock that has the potential to have more discrete observations. Such trading frequency would also affect the distribution properties of stocks. Let us consider the general distribution properties of orders placed for stocks and also the returns generated through trading of two stocks,

namely, Telstra Ltd. and Ten Network Ltd. for January and February, 2012. The stocks are mainly different in terms of their size. Telstra is a large stock with higher trading frequency, whereas Ten Network Ltd. is relatively small stock with lower trading frequency. Figure 3.2(a) and 3.2(b) display the orders distribution of Telstra Ltd. and Ten Network Ltd. stocks. Both stocks have fairly normal distribution with mean OIB of around 22,000 excess buy orders and 870 excess sell orders for Telstra and Ten, respectively. Figures 3.2(c) and 3.2(d), display the OIB of both stocks by randomly deleting 10% of the orders. This exercise helps us to visualize order book with missing data.

Given that order arrival is random, hidden orders are expected to be distributed randomly and hence randomly missing few observations (that are visible) represent hidden order book for distribution purposes. As shown in the figures, the distributional properties do not change with missing data and the mean values are similar to the full OIB data. We repeat this exercise by missing 20% of the observations. As shown in Figures 3.2(e) and 3.2(f), the distributional properties are not affected by randomly missing even 20% of orders.

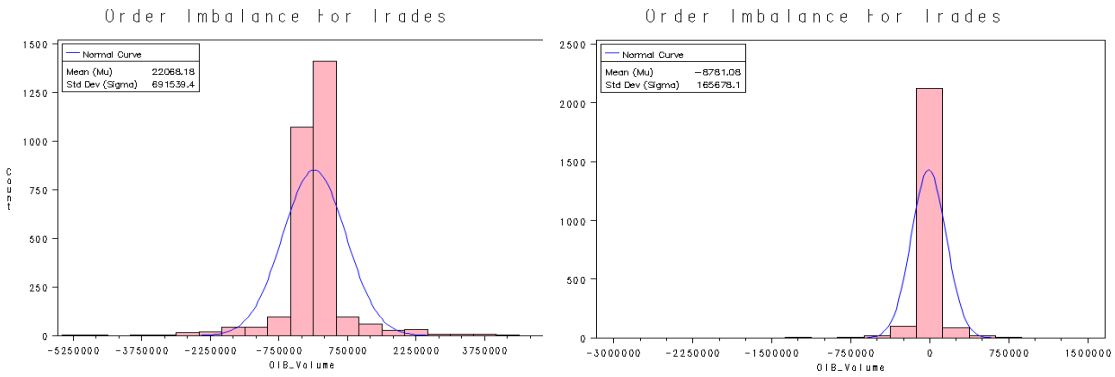


Figure 3.2 (a) and 3.2 (b) Distribution of Telstra and Ten Network

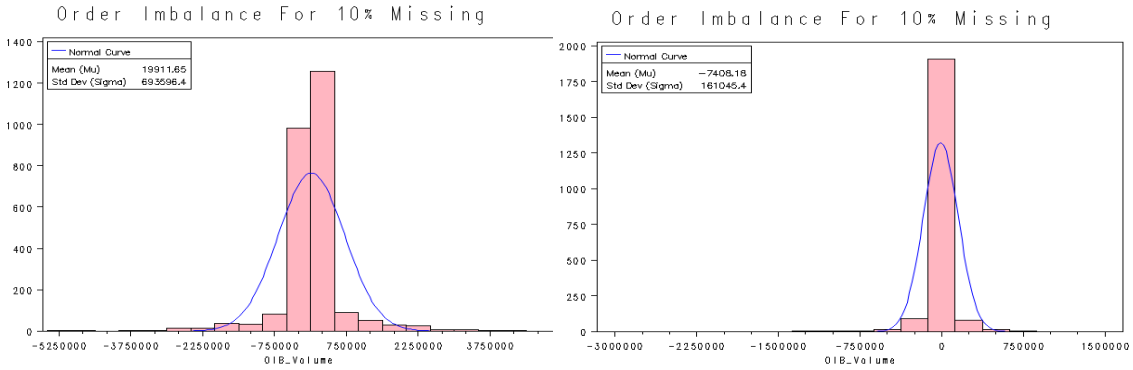


Figure 3.2 (c) and 3.2 (d) Distribution of Telstra and Ten Network with 10% Missing Data

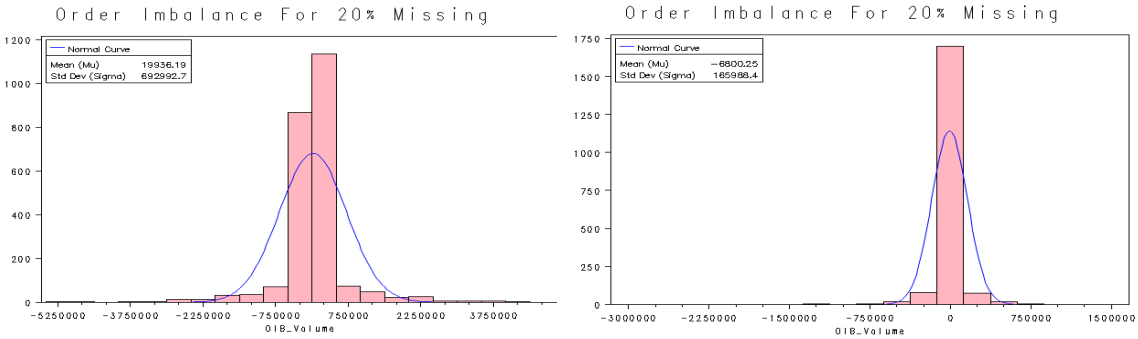


Figure 3.2 (e) and 3.2 (f) Distribution of Telstra and Ten Network with 20% Missing Data

These figures provide strong basis for two important assumptions: that orders have normal distribution and that randomly missing data does not affect the distributional properties of the orders. However, it is important to note that large orders due to private information traders and institutional traders can potentially lead to fat tails in the distribution and thus violating the normal distribution assumption of this example.

Continuing the discussion on maximum likelihood estimation for missing data, the dataset in Table 3.2 illustrates the idea of missingness. This data are designed to mimic order imbalance from complete orders in OIB column and order imbalance calculated from Trade book in TOIB column which represents MAR data. Some of the data from

TOIB column are missed randomly in order to mimic a MCAR mechanism and shown in the third column as MTOIB which has Missing TOIB values. This data set is too small for a serious application of the maximum likelihood estimation, but is useful for illustrating the mechanics of the procedure.

3.5.2.2 Log Likelihood for Missing Data

The starting point for a maximum likelihood analysis is to specify a distribution for the population data. The mathematical machinery behind maximum likelihood relies on a probability density function that describes the shape of the multivariate normal distribution. Substituting a score vector and a set of population parameter values into the density function returns a likelihood value that quantifies the relative probability of drawing the scores from a normally distributed population. Because likelihood values tend to be very small numbers that are prone to rounding error, it is more typical to work with the natural logarithm of the likelihood values (i.e., the log-likelihood).

Repeating the equation 3.4, the log likelihood for complete data is

$$\log L_i = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)$$

where *k* is the number of variables, Y_i is the score vector for case *i*, and μ and $\boldsymbol{\Sigma}$ are the population mean vector and covariance matrix, respectively. The key portion of the formula is the Mahalnobis distance value, $(Y_i - \mu)^T \Sigma^{-1} (Y_i - \mu)$. Mahalnobis distance is a squared *z* score that quantifies the standardized distance between an individual's data points and the center of the multivariate normal distribution.

With missing data, the log-likelihood for case i is

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(Y_i - \mu_i)^T \Sigma^{-1} (Y_i - \mu_i) \quad (3.5)$$

where k_i is the number of complete data points for that case and the remaining terms have the same meaning as they did in equation 3.4. At first glance, the two log-likelihood formulae look identical, except for the fact that the missing data log-likelihood has an *i* subscript next to the parameter matrices. This subscript is important and denotes the possibility that the size and the contents of the matrices can vary across individual observations, such that the log-likelihood computations for case *i* depend only on the variables and the parameters for which that case has complete data.

To illustrate the missing data log-likelihood, we consider the data in Table 3.2 to estimate the mean vector and the covariance matrix. Estimating these parameters is relatively straight forward with the complete data but requires an iterative optimization algorithm when some of the data are missing. The iterative algorithm is discussed in detail in Chapter 4.

For the sake of demonstration, suppose that the population parameters at a particular iteration are as follows:

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{MTOIB} \\ \hat{\mu}_{Return} \end{bmatrix} = \begin{bmatrix} 513 \\ 287 \\ 0.584 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\sigma}^2_{OIB} & \hat{\sigma}_{OIB,MTOIB} & \hat{\sigma}_{OIB,Return} \\ \hat{\sigma}_{MTOIB,OIB} & \hat{\sigma}^2_{MTOIB} & \hat{\sigma}_{MTOIB,Return} \\ \hat{\sigma}_{Return,OIB} & \hat{\sigma}_{Return,MTOIB} & \hat{\sigma}^2_{Return} \end{bmatrix} = \begin{bmatrix} 32277.69 & -945.75 & 0.51 \\ -945.75 & 18754 & 0.35 \\ 0.51 & 0.35 & 0.01 \end{bmatrix}$$

The log-likelihood computations for each individual depend only on the variables and the parameters for which a case has complete data. This implies that the log-

likelihood formula looks slightly different for each missing data pattern. The dataset in Table 3.2 has two missing data patterns: 1) cases with OIB and MTOIB data and 2) cases with complete data on all three variables.

To begin with, we consider the case 2, where all data are available. Consider the score 692 for complete OIB from Order book and 346 for missing MTOIB and 0.59543 for Return in the Table 3.2. Because this individual has complete data, the log-likelihood computations involve every element in the mean vector and the covariance matrix, as follows:

$$\log L_i$$

$$= -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} \hat{\sigma}^2_{OIB} & \hat{\sigma}_{OIB,MTOIB} & \hat{\sigma}_{OIB,Return} \\ \hat{\sigma}_{MTOIB,OIB} & \hat{\sigma}^2_{MTOIB} & \hat{\sigma}_{MTOIB,Return} \\ \hat{\sigma}_{Return,OIB} & \hat{\sigma}_{Return,MTOIB} & \hat{\sigma}^2_{Return} \end{vmatrix} - \frac{1}{2}\left(\begin{bmatrix} OIB_i \\ MTOIB_i \\ Return_i \end{bmatrix}\right.$$

$$\left. - \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{MTOIB} \\ \hat{\mu}_{Return} \end{bmatrix}\right)^T \Sigma \begin{vmatrix} \hat{\sigma}^2_{OIB} & \hat{\sigma}_{OIB,MTOIB} & \hat{\sigma}_{OIB,Return} \\ \hat{\sigma}_{MTOIB,OIB} & \hat{\sigma}^2_{MTOIB} & \hat{\sigma}_{MTOIB,Return} \\ \hat{\sigma}_{Return,OIB} & \hat{\sigma}_{Return,MTOIB} & \hat{\sigma}^2_{Return} \end{vmatrix}^{-1} \left(\begin{bmatrix} OIB_i \\ MTOIB_i \\ Return_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{MTOIB} \\ \hat{\mu}_{Return} \end{bmatrix}\right)$$

$$=$$

$$-\frac{3}{2}\log(2\pi) -$$

$$\frac{1}{2}\log\begin{vmatrix} 32277.69 & -945.75 & 0.51 \\ -945.75 & 18754 & 0.35 \\ 0.51 & 0.35 & 0.01 \end{vmatrix} - \frac{1}{2}\left(\begin{bmatrix} 692 \\ 346 \\ 0.59543 \end{bmatrix} -\right.$$

$$\left.\begin{bmatrix} 513 \\ 287 \\ 0.584 \end{bmatrix}\right)^T \Sigma \begin{vmatrix} 32277.69 & -945.75 & 0.51 \\ -945.75 & 18754 & 0.35 \\ 0.51 & 0.35 & 0.01 \end{vmatrix}^{-1} \left(\begin{bmatrix} 692 \\ 346 \\ 0.59543 \end{bmatrix} - \begin{bmatrix} 513 \\ 287 \\ 0.584 \end{bmatrix}\right) = -10.48$$

The log-likelihood computations for the remaining complete cases follows the same procedure, but different score values.

Next, consider the subsample with OIB and Return scores. These individuals have missing values in the column, MTOIB, so it is no longer possible to use all three variables to compute log likelihood. The missing data log-likelihood accommodates this situation by ignoring the parameters that correspond to the missing MTOIB values. For example, consider the individual with OIB and $Return$ scores of 884 and 0.590361 respectively. Eliminating the MTOIB from the mean vector and the covariance matrix leaves the following subset of parameter estimates.

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{Return} \end{bmatrix} = \begin{bmatrix} 513 \\ 0.584 \end{bmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{i}} = \begin{bmatrix} \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\boldsymbol{OIB}} & \hat{\boldsymbol{\sigma}}_{\boldsymbol{OIB},Return} \\ \hat{\boldsymbol{\sigma}}_{Return,\boldsymbol{OIB}} & \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{Return} \end{bmatrix} = \begin{bmatrix} 32277.69 & 0.51 \\ 0.51 & 0.01 \end{bmatrix}$$

The log-likelihood computations use only these parameter values, as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\left|\begin{matrix} \hat{\sigma}^2_{OIB} & \hat{\sigma}_{OIB,Return} \\ \hat{\sigma}_{Return,OIB} & \hat{\sigma}^2_{Return} \end{matrix}\right| - \frac{1}{2}\left(\begin{bmatrix} OIB_i \\ Return_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{Return} \end{bmatrix}\right)^T \Sigma \left|\begin{matrix} \hat{\sigma}^2_{OIB} & \hat{\sigma}_{OIB,Return} \\ \hat{\sigma}_{Return,OIB} & \hat{\sigma}^2_{Return} \end{matrix}\right|^{-1} \left(\begin{bmatrix} OIB_i \\ Return_i \end{bmatrix} - \begin{bmatrix} \hat{\mu}_{OIB} \\ \hat{\mu}_{Return} \end{bmatrix}\right)$$

$$= -\frac{2}{2}\log(2\pi) - \frac{1}{2}\log\left|\begin{matrix} 32277.69 & 0.51 \\ 0.51 & 0.01 \end{matrix}\right| - \frac{1}{2}\left(\begin{bmatrix} 884 \\ 0.59543 \end{bmatrix} - \begin{bmatrix} 513 \\ 0.584 \end{bmatrix}\right)^T \Sigma \left|\begin{matrix} 32277.69 & 0.51 \\ 0.51 & 0.01 \end{matrix}\right|^{-1} \left(\begin{bmatrix} 884 \\ 0.59543 \end{bmatrix} - \begin{bmatrix} 513 \\ 0.584 \end{bmatrix}\right) = -4.64$$

Notice that the log-likelihood equation no longer contains any reference to the MTOIB variable. Thus, the resulting log-likelihood value is relative probability of drawing two scores from a bivariate normal distribution with a mean vector and

covariance matrix equal to $\hat{\mu}_i$ and $\hat{\boldsymbol{\Sigma}}_i$, respectively. The log-likelihood computations for cases that follow the same missing data pattern follow the same method.

Likewise, log-likelihood values for each individual case for complete data and missing data are computed and the sample log-likelihood values are calculated by summing these log-likelihood values. Despite the missing values, the sample log-likelihood is still summary measure that quantifies the joint probability of drawing the observed data from a normally distributed population with a particular mean vector and covariance matrix.

Conceptually, an iterative optimization algorithm repeats the log-likelihood computations many times, each time with different estimates of the population parameters. Each unique combination of parameter estimates yields a different log-likelihood value. The goal of estimation is to identify the particular constellation of estimates that produce the highest log-likelihood and thus the best fit to the data. Importantly, the estimation algorithm does not need to impute or replace the missing values. Rather, it uses all of the available data to estimate the parameters and the standard errors.

3.6 Prediction from Missing Data

Following the discussion in Section 3.4, it is not necessarily obvious why including the incomplete data records improves the accuracy of the resulting parameter estimates. This section provides deeper insights into this estimation process related issues by using a bivariate analysis. One of the common methods discussed in the Section 3.4 is the list-wise deletion method. In comparison to maximum likelihood estimates of the missing

dataset, list-wise deletion excludes cases from the lower tail[24] of the MTOIB distribution. Consequently, the list-wise deletion's mean estimates will be higher. In contrast, the maximum likelihood estimates are relatively similar to the complete data. These estimates are consistent with Rubin's (1996) theoretical predictions for a MAR mechanism. Table 3.3 shows two columns, one with complete data of order imbalance from order book and other with missing data of order imbalance from trade book.

Table 3.3 Order Imbalance Dataset with Missing Data

OIB (Orders)	MTOIB
807	272
692	346
884	----
385	271
623	----
635	229
461	201
340	136
464	----
337	176
519	243

With missing data, the individual log-likelihood computations depend only on the variables and the parameter estimates for which a case has data. Let us consider a bivariate analysis to illustrate how the maximum likelihood estimates for missing data are relatively similar to the complete data. A bivariate analysis has just two missing data patterns (i.e., cases with complete data (both OIB and MTOIB) and cases with OIB only as shown in the table below. There are two log-likelihood formulas for the two cases.

[24] Ends of the distribution curve.

The individual log-likelihood equation for the subsample of the order imbalance with complete data is

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log\begin{vmatrix} \sigma^2{}_{OIB} & \sigma_{OIB,MTOIB} \\ \sigma_{MTOIB,OIB} & \sigma^2{}_{MTOIB} \end{vmatrix} - \frac{1}{2}\left(\begin{bmatrix} OIB_i \\ MTOIB_i \end{bmatrix} - \begin{bmatrix} \mu_{OIB} \\ \mu_{MTOIB} \end{bmatrix} \right)^T \Sigma \begin{vmatrix} \sigma^2{}_{OIB} & \sigma_{OIB,MTOIB} \\ \sigma_{MTOIB,OIB} & \sigma^2{}_{MTOIB} \end{vmatrix}^{-1} \left(\begin{bmatrix} OIB_i \\ MTOIB_i \end{bmatrix} - \begin{bmatrix} \mu_{OIB} \\ \mu_{MTOIB} \end{bmatrix} \right) \quad (3.6)$$

and eliminating MTOIB parameters gives the individual log-likelihood equation with incomplete data, as follows:

$$\log L_i = -\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2{}_{OIB}| - \frac{(OIB_i - \mu_{OIB})^2}{2\sigma^2{}_{OIB}} \quad (3.7)$$

Finally, summing the previous equations across the entire sample gives the sample log-likelihood

$$\log L = \{-n_C(\frac{k_i}{2}\log[2\pi] - \frac{1}{2}\log|\Sigma|) - \frac{1}{2}\sum_{i=1}^{n_C}(Y_i - \mu)^T\Sigma^{-1}(Y_i - \mu)\}$$

$$-n_M(\frac{k_i}{2}\log(2\pi) - \frac{1}{2}\log|\sigma^2{}_{OIB}| - \frac{1}{2\sigma^2{}_{OIB}}\sum_{i=1}^{n_M}(OIB_i - \mu_{OIB})^2 \quad (3.8)$$

$$= \{\log L_{Complete}\} + \{\log L_{Incomplete}\}$$

Where n_C is the number of complete cases, and n_M is the number of the incomplete cases. Equation 3.8 is useful because it partitions the sample log-likelihood into two components. The bracketed terms reflect the contribution of complete cases to the sample log-likelihood, and the remaining terms contain the additional information from the incomplete data records. The portion of the log-likelihood equation for the

incomplete cases serves as a correction factor that steers the estimator to a more accurate set of parameter estimates.

In what follows, this intuition is applied while building the Expectations Maximization algorithm in Chapter four, where Equation 3.8 maximizes the likelihoods through an iterative process. This process augments the total likelihood estimated with additional information, supplied by incomplete cases in each iteration.

To understand this clearly, let us look at the Table 3.4 that has some hypothetical values for OIB and MTOIB and compares the maximum likelihood estimate with list-wise deletion method to explain clearly how missing data borrows the information from complete data. To illustrate how the incomplete data records affect estimation, Table 3.4 shows the sample log-likelihood for different combinations of the OIB and MTOIB means. For simplicity, OIB estimates are limited to 513.00 and 550 (these are maximum likelihood and list-wise deletion estimates, respectively).The column labeled *Log L_Complete* contains the sample log-likelihood values from list-wise deletion analysis (i.e., maximum likelihood estimation based only on the bracketed terms in equation 3.8; this does not contain missing data); the column labeled *Log L_Incomplete* shows the log-likelihood contribution for the incomplete data records; and the *LogL* column gives the sample log-likelihood values for maximum likelihood missing data handling (i.e., the sum of $\log L_{Complete}$ and $\log L_{Incomplete}$). As seen in $\log L_{Complete}$ column, a list-wise deletion analysis would produce estimates of μ_{OIB} = 550 and μ_{MTOIB}= 285 because this combination of parameter values has the highest (i.e., least negative; marked in yellow) log-likelihood value.

Table 3.4 Sample Log-Likelihood Values for different Combinations of the OIB and MTOIB Means

		Log- Likelihood		
μ OIB	**μ TOIB = 268**	**Log L_Complete**	**Log L_InComplete**	**Log L**
513.00	250.00	-140.40	-53.03	-193.43
	255.00	-140.35	-53.03	-193.38
	260.00	-140.31	-53.03	-193.34
	265.00	-140.29	-53.03	-193.32
	270.00	-140.29	-53.03	-193.32
	275.00	-140.30	-53.03	-193.33
	280.00	-140.34	-53.03	-193.37
	285.00	-140.39	-53.03	-193.42
	290.00	-140.46	-53.03	-193.49
	290.00	-140.49	-53.03	-193.52
	μ MTOIB = 287			
550.00	250.00	-132.71	-72.21	-204.92
	255.00	-132.64	-72.21	-204.85
	260.00	-132.57	-72.21	-204.78
	265.00	-132.52	-72.21	-204.73
	270.00	-132.48	-72.21	-204.69
	275.00	-132.45	-72.21	-204.66
	280.00	-132.43	-72.21	-204.64
	285.00	-132.42	-72.21	-204.63
	290.00	-132.43	-72.21	-204.64
	295.00	-132.44	-72.21	-204.65

Next, the log $L_{Incomplete}$ column gives the contribution of the 10 incomplete cases to the sample log-likelihood. Because these do not have MTOIB values, the log-likelihood values are constant across different estimates of the MTOIB mean (i.e., Equation 3.7 depends only on OIB parameters). However, the incomplete data records do carry information about the OIB mean, and the log-likelihood values suggest that μ_{OIB} = 513.00 is more plausible than μ_{OIB} = 550.00 (i.e., the log-likelihood for μ_{OIB} = 513.00 is higher than that of μ_{OIB} = 550.00). Finally, the Log L column gives the sample

log-likelihood values for the maximum likelihood missing data handling. As can be seen, μ_{OIB} = 513.00 and μ_{TOIB} = 270 (in yellow) provide the best fit to the data because this combination of parameter values has the highest log-likelihood (-193.32).

Towards the end of Section 3.4.1, an expectation: "*High likelihood of the estimates to the population mean can be obtained when the observed mean is close to the population mean"* is set. By the above explanation that, the mean value, μ_{TOIB} = 268 (between the values of μ_{TOIB} = 265 and μ_{TOIB} = 270, in yellow), with μ_{OIB} = 513 provide the best fit as the log likelihood value here is the highest (low negative). By this it can be reiterated that the expectation set in Section 3.4.1 is proved.

Mathematically, the goal of maximum likelihood estimation is to identify the parameter values that minimize the standardized distances between the data points and the center of a multivariate normal distribution. Whenever the estimation process involves a set of model parameters, fine tuning one estimate can lead to changes in the other estimates. This is precisely what happened in the bivariate analysis example. Specifically, the log-likelihood values in the $\log L_{Incomplete}$ column strongly favor a lower value for the OIB mean as it includes missing values as well. Including these incomplete data records in the analysis therefore pulls the OIB mean down to a value that is identical to that of complete data. Higher values for MTOIB mean are an unlikely match for an OIB mean, so the downward adjustment to the OIB average effectively steers the estimator toward a MTOIB mean that more closely matches that of the complete data.

In effect, maximum likelihood estimation improves the accuracy of the parameter estimates by "borrowing" information from the observed data (i.e., the OIB

scores), some of which is contained in the incomplete data records. This analysis justifies using maximum likelihood estimation procedure to improve predictions when data are incomplete or missing. In the case of this thesis, maximum likelihood estimation helps in borrowing information from the incomplete records of the observed orders and improves the estimation accuracy.

3.7 Summary

The goal of this thesis is to recover missing information of unobservable features related to stock orders that have significant influence on price formation in stock markets and thereby enhance return prediction. Given that the observable orders based OIB calculation from trade book result in underestimation of actual OIB level, a maximum likelihood estimation process is proposed in the thesis.

After justifying that underestimation or the missingness of the data follows MAR classification, using the maximum likelihood estimation, different combinations of population parameter values (mean and variance) are auditioned until the particular constellation of values that produce highest log-likelihood value are identified. This chapter provides the foundation for understanding the mechanics of the proposed estimation procedure. It is shown that the log likelihood values for incomplete data (OIB from trade book) favors a lower value for OIB mean. This downward adjustment to the OIB average effectively steers the estimator toward a MTIOB (OIB mean for missing data) that is identical to that of complete data. This is done by “borrowing” information from the observed data, some of which is contained in the incomplete data records and improves prediction accuracy.

Using real data of a sample stock, AGK (AGL Energy Ltd.), this chapter demonstrates that estimating by using incomplete records, as against deleting missing information, improves the accuracy of the parameter estimates. Backed by this theoretical and empirical justification, the next chapter devises a methodology for predicting stock returns using the trade book data.

Chapter 4

PREDICTION METHOD THROUGH ORDER BOOK MAPPING

4.1 Introduction

In order to exploit the stock market arbitrage, one needs to make an accurate prediction of the future price movements of stocks. As argued in this thesis, complete order book information is a key determinant of the short run temporal changes in the stock prices. In Chapter 3, in the example from Table 3.1, it is noted that there is a disparity between the order imbalance calculated from complete data and TOIB (trade book's OIB). Hence, using observable trade book data leads to underestimation problem. Chapter 3 proposes a maximum likelihood estimation (MLE) procedure to overcome such underestimation problem. Both the theoretical and empirical justification provides the foundation to devise a systematic methodology for predicting stock returns using the observable trade book data.

In this chapter, using Chapter 3 as the base, a detailed method of implementing the proposed theory is explored. Chapter 3 mainly focuses on the methodology of how we can improve the estimation with missing data. In other words, how order imbalance can be accurately estimated from an incomplete dataset of order imbalance. In this chapter, this logic gets extended to propose a complete process to predict future returns in the presence of missing data. This involves establishing a functional relation between order

imbalance and future returns, addressing the multi dimensionality problem with complexity associated with multiple parameters, establishing joint probability functions between orders and returns, and discussing robust procedures to improve return prediction accuracy.

Guided by the financial economic theory, we establish a functional relation between orders and returns in Section 4.2, with a proposed framework by defining the probabilistic distributional properties of all parameters. The theoretical idea proposed in Chapter 3 is decomposed in this chapter into the actual application, through systematic estimation procedure in this section. Next, in Section 4.3, we introduce Relational Markov Network (RMN) as a procedure to map the matched orders and their corresponding interactions (returns). Here, we first discuss the problems associated with regression models and suggest a relational framework to overcome these by representing orders, order imbalance, order characteristics and returns as a relational model. However, calculating conditional probabilities of each order interaction (return) instance is highly intractable in complex relational network. To overcome this, we propose a simplification process using Markov Chain Monte Carlo method. In Section 4.4, we present the step-wise algorithmic procedure.

In Section 4.5, we address how the complexities of the maximum likelihood process are being handled through established machine learning procedures. We present the probabilistic model to establish the distributional properties of order characteristics and their interactions by defining their theoretical distributions and also their joint probability distribution function to estimate the mean of the distribution.

It is argued in Chapter 3 that estimation accuracy can be improved by applying maximum likelihood procedure by incorporating missingness as part of the estimation procedure. Complex applications of maximum likelihood estimation generally require iterative optimization algorithms, such as Expectations Maximization Algorithm as an integrated learning method to estimate the missing orders. This is discussed in detail in the Section 4.6.

4.2 Estimation Procedure for Future Returns Prediction using OIB Estimates

Order book data are quite complex with several relations among orders placed by trades. In addition to that, we have to establish a relation between orders in the order book with stock prices. These prices are influenced by orders and change due to change in the orders. For joint estimation of all the variables, we need a combination of methods and procedures that can deal with the multi dimensionality issues associated with several variables and also to reduce overall estimation errors leading to improved accuracy in predicting returns. Figure 4.1 depicts the methodological framework. As shown in the figure, we first establish the relation between orders and returns. Later, we propose the steps involved in the estimation procedure.

Framework Steps

The framework steps described below lay foundation for the empirical estimation process. The main objective of these steps is to decompose the complex estimation process for tractability. First step is an extension of the probabilistic theory proposed in Chapter 3, in the context of financial theory. Given that the thesis aims to establish the

predictive role of order imbalance on future returns, in the first step, the logic behind the functional relation between the two variables needs to be justified. As a second step, the probabilistic theory proposed in Chapter 3 is extended to incorporate returns into the joint estimation procedure.

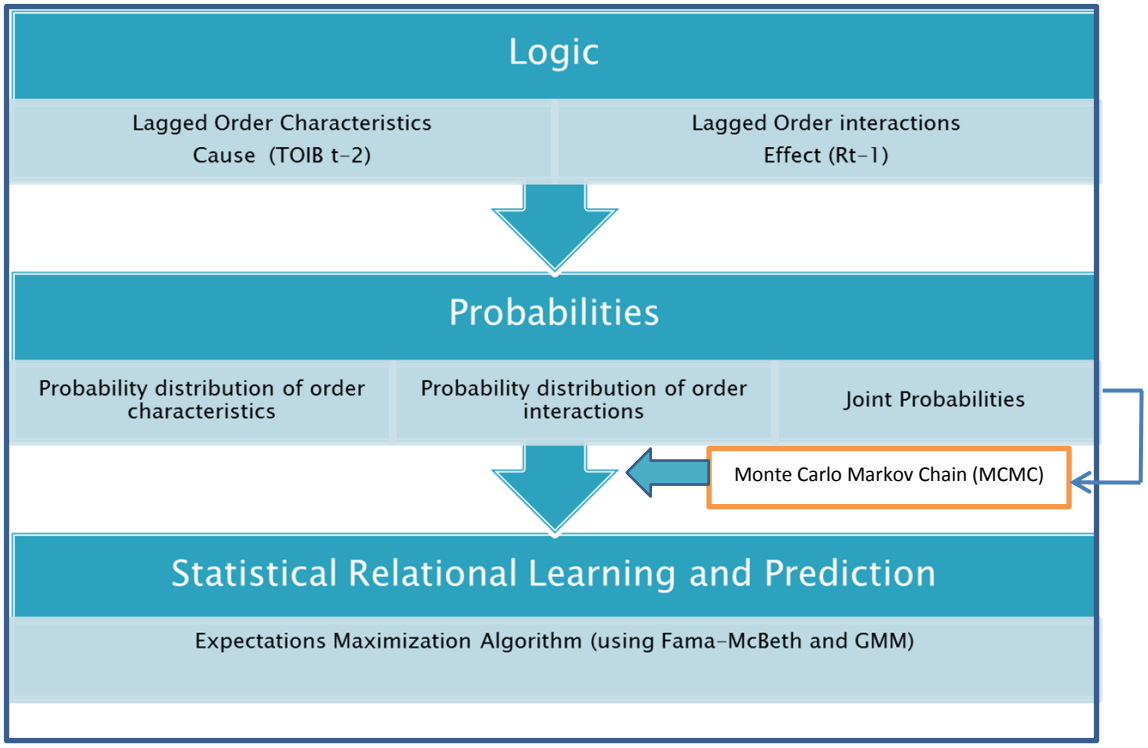


Figure 4.1 Proposed Framework

The complexities such as multidimensionality, cross sectional correlation between variables that arise due to inter related multiple relationships between variables and their interactions are dealt in the third and fourth steps.

Step 1 - Logic: The study of how orders and their interactions influence price changes is the central theme of the thesis. This influence, in order to model as a predictive relationship, should have a clear cause and effect relationship. This is the logic step of the methodology. Financial economics literature has established both

theoretical and empirical justification that orders cause temporal price changes of the corresponding stock (Chordia et al., 2005). Return at time *t*, is caused by order imbalance (TOIB) at time *t-1* (lagged) is the logic employed in the first step. As described later, we establish this relational logic through a relational network model in this chapter.

Step 2 - Probabilities: In this step, the probability distribution for order characteristics, order interaction, and their joint probability distribution are calculated. A joint probabilistic model is needed for the entire collection of orders and order interactions. This will help demonstrate that the distribution of unobservable orders can be obtained through machine learning to complete the estimate of the collection of orders and their interactions. We propose to use the Relational Markov Networks (RMN) framework, a probabilistic model discussed later in this chapter, for estimating the joint distribution of orders (as discussed in chapter 3). We justify its usage by highlighting the problems associated with regression models while dealing with relational datasets like ours.

Step 3 - MCMC Method: The joint distribution as calculated in Step 2 suffers from multi dimensionality issues. Multi dimensionality issues arise due to intractable relationships that are needed to estimate conditional probabilities of order interactions. In order to simplify this complexity, we apply Monte Carlo Markov Chain (MCMC) method (described in detail in the following section, when explaining how relational Markov network is implemented). MCMC integrates and approximates unknown parameters.

Step 4 - Statistical Relational Learning and Prediction: After the densities of the variables are estimated using MCMC, we then implement statistical relational learning and prediction. Here, the Expectation Maximisation algorithm is used to learn the missing orders' order imbalance estimates. We use the Fama-Macbeth (Fama and Macbeth, 1973) procedure as it corrects standard errors for cross sectional correlation between variables.

For this procedure, the first step calculates firm level time series beta[25] estimates and the second step runs cross-sectional regression by pooling individual time series estimates. There is a disadvantage in this procedure. It does not correct time series correlation (auto correlation). Given that we use pooled sample of several stocks, there is every chance that the attributes of one stock are spuriously correlated with other stock. This leads to biased estimates. This is generally called as cross-sectional correlation problem. Likewise, within one stock, over time, its returns might be correlated with its own past returns and hence such auto correlation leads to biased estimates. This is generally called time-series correlation. Both cross-sectional and time-series correlations violate the basic underlying assumptions of regression methodology. In order to overcome this issue, we use the Generalised Methods of Moments (GMM) estimation procedure to correct the time series correlation. These methods and their context specific explanations are discussed in detail in Chapter 5.

As described in the first step, we first need to establish probability distribution functions for estimating and predicting future returns. Given that order imbalance and returns are interrelated, we have to define the relational network between variables and

[25] Beta is the coefficient of a given determinant (for instance, the estimate of order imbalance).

later the joint probabilistic distribution functions of the variables. The next section describes the Relational Markov Network (RMN) framework and how it overcomes the estimation problem associated with the existing literature (Section 4.3.2) and thus how it can be applied to the research problem (Section 4.3.3).

4.3 Relational Markov Networks (RMN)

In Chapter 3, it is established that missing data problem can be addressed by maximum likelihood procedure that incorporates missing data through imputation process. Such procedure will improve the estimates of the datasets that suffers from missing data problem. However, it does not ensure better prediction of future returns. This is due to interaction of several variables in the estimation procedure. It is important to establish relationships between variables in the overall estimation process.

4.3.1 The Problems with Regression Models

Existing studies in finance, in particular, order imbalance based prediction studies, namely Chordia et al. (2009) use standard regression models for predicting the effect of order imbalance on price movements or returns that are caused due to order interactions. The main issue with this approach is that orders, probability of their interactions and the corresponding price movements are correlated. This violates the underlying assumptions[26] of regression models leading to biased estimates. For instance, if the current ask is $10 (SO1 with an order size of 1000 shares) and the bid is at $9 (BO1 with an order size of 2000 shares). A new bid (BO2) placed at $ 10 (for order interaction or

[26] Variables should not be correlated. Correlations between variables violate the independent and identical distribution (i.i.d) assumption.

matching) is influenced by the bid price of $9 and also on the positive order imbalance of 1000 shares (2000-1000). This implies that the current order interaction (SO1 and BO2 matching of Bid and Ask at $10) at time *t,* R_t, the return is influenced by the current orders at time *t,* also the past orders at time *t-1* and their corresponding order imbalance at time *t-1*, OIB_{t-1}.

Let us extend this example to one more layer of complexity. Let us assume that there are several orders behind the ask order of 1000 shares with a price higher than $10. Likewise, there is similar piling of orders on the bid side. These several orders, although do not interact, nevertheless, contain some probability of interaction. Their joint probability can influence the order interaction that may not directly be connected to the matched orders. One problem of not accounting for such complex network based information while predicting price movements is that, we may underestimate the predictive ability of order imbalance. Such complex web of inter-relationships due to sharing of same orders in different OIB templates (at temporal level) leads to limitations arising due to correlations between the variables of interest, while estimating, using regression models.

This implies that one need to consider the entire network of orders, along with the interacted orders to produce joint prediction. Hence, there is a need to develop an explicitly modeled theoretical framework that can show all such dependencies and create a unified probabilistic model for encoding such reasoning and for effective learning of the order interaction network. As shown in Figure 4.2, if we assume two OIB levels, namely, V1 and V2, at two consecutive time intervals, t and t-1, respectively, we can

show that they both are related in terms of orders in the order imbalance templates of V1 and V2. However, they are different only in terms of orders BO 2 and SO 3 (that exist only in V2). The solid arrowed lines show the match between orders and the dotted arrow shows the dependency of a particular unmatched order that is carried forward to the next time interval, causing the interdependency between order imbalance templates.

For instance, the interdependency between templates V1 and V2 is due to sharing of BO 1, SO 1 and SO 2 and hence becoming part of a relational network. In addition to that, arrival of a new order in template V2 (BO 2), leads to order interaction as stipulated by the matching algorithm of the stock exchange.

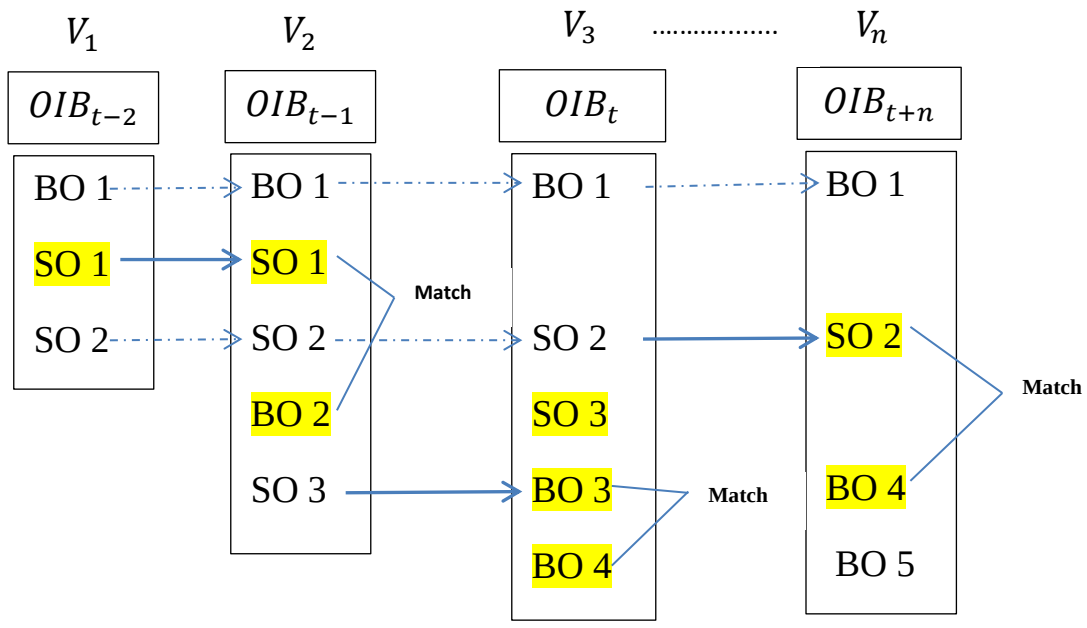


Figure 4.2 Influences of Current and Previous Orders on Order Matching

In this case, we observe only matched orders and the corresponding return $\boldsymbol{R_{t-1}}$ (as shown in Figure 4.3 which is similar to Figure 4.2) and the rest of the information of

the relational network is unobservable. Likewise, if we consider *n* consecutive OIB levels, separated by *n* time intervals, they are interdependent and have complex network connections.

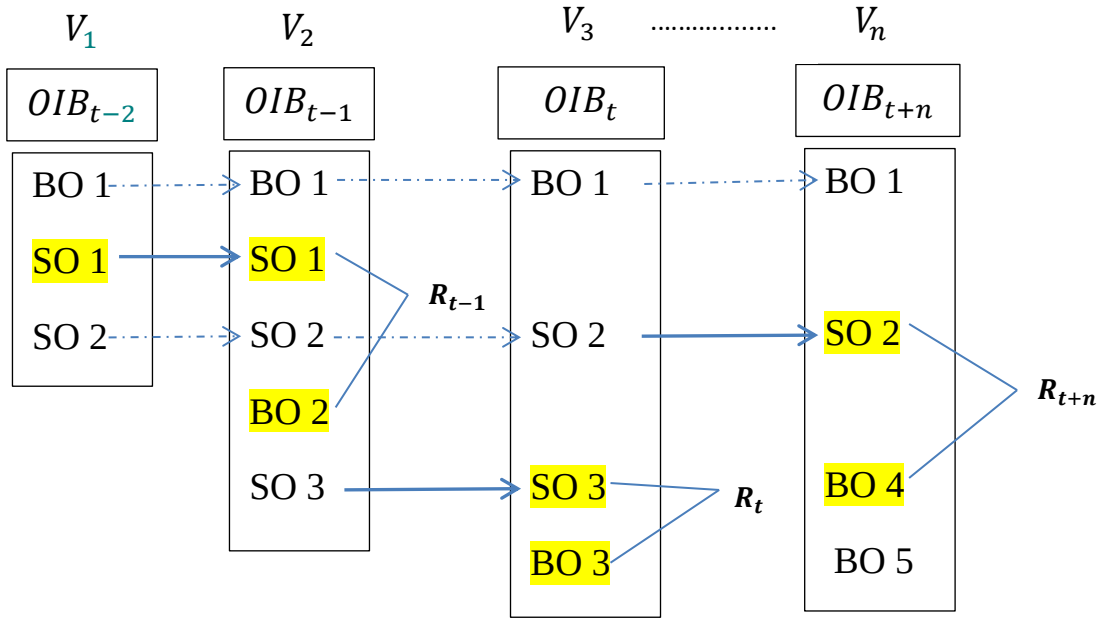


Figure 4.3 OIB Levels and their Relationships

Ignoring all this information of the relational network, leads to underestimation bias. Chordia et al. (2005) use lag return and lag OIB as the determinants of contemporaneous (current) returns. They use Ordinary Least Squares (OLS) regression models to estimate the relation between contemporaneous return (R_t), lag returns (R_{t-1}) and lag order imbalance (OIB_{t-1}) as in Equation 4.1.

In other words, they test the hypothesis whether future returns depend on the past returns and past order imbalance. However, there is a major issue with this estimation procedure as it does not define probabilistic dependencies between various orders at different order imbalance levels. Figure 4.3 explains a relational probabilistic network

model, to explicitly define these dependencies. Figure 4.4 represents the determinants of returns established in finance literature (Chordia et al., 2005). The Equation 4.1 shows that current return MR_t is determined by the past return, MR_{t-1} and past order imbalance level OIB_{t-1}. α_0 is the intercept, α_1, α_2, α_3 are the coefficient terms and ε_t is the error term. This relationship is depicted in Figure 4.4 (a).

$$MR_t = \alpha_0 + \alpha_1 MR_{t-1} + \alpha_2 OIB_{t-1} + \varepsilon_t \quad (4.1)$$

However, as shown in the Figure 4.4 (b), some of the orders, those influence both contemporaneous and lag returns, are part of lagged order imbalance.

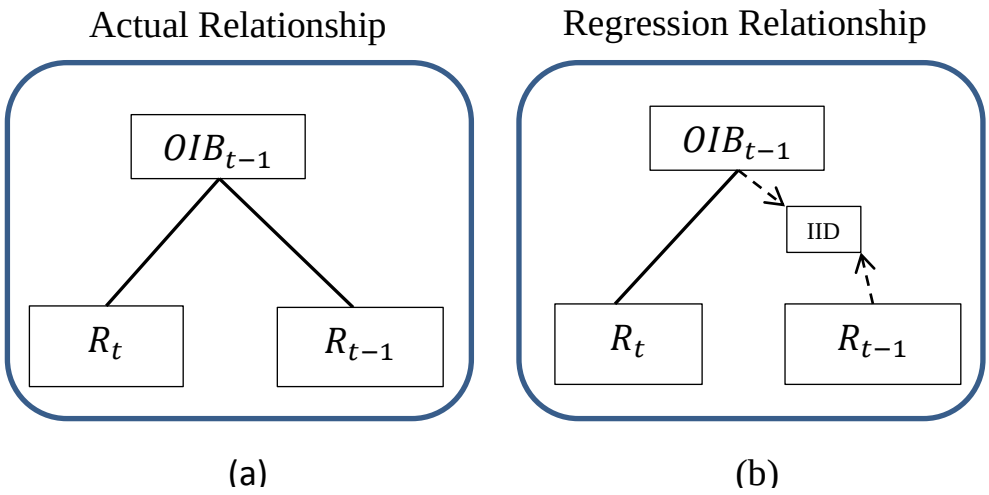


Figure 4.4 Relationship Comparison

In other words, orders that are part of OIB_{t-1} are carried forward to OIB_t period. This is due to the orders that are not matched, continue to remain and they still continue to influence the returns in the next period. In other words, R_t is not influenced by R_{t-1} and OIB_{t-1} alone. It is influenced by OIB_{t-2} and the OIBs of previous intervals where the unmatched orders are carried forward. So, OIB_{t-1} is not independent. It has a

relationship with all those orders of previous periods that are carried forward. This gives rise to multiple relationships between variables. This violates the independent and identical distribution (i.i.d) assumption imposed on the independent variables OIB_{t-1} and R_{t-1} in regression framework.

Machine learning methods propose graphical models to establish relationship between variables for estimating the joint probabilistic functions of the collective distribution. Conditional Random Fields (CRFs), which are undirected graphical models, are widely used in this context. CRFs are discriminative models that have been shown to out-perform generative approaches such as Hidden Markov Models and Markov Random Fields in areas such as natural language processing (Lafferty et al., 2001). Relational Markov Networks (RMNs) extend CRFs by providing a relational language for describing clique structures and enforcing parameter sharing at the template level. Thereby RMNs are an extremely flexible and concise framework for defining features.

Researchers used RMN framework extensively in classifying problems, especially for information retrieval. Rather than classifying each entity separately, a form of collective classification, where class labels of all of the entities are together, and thereby one can take advantage of the correlations between the several related entities. Also, RMNs can help to predict the entities that are related to others with multiple relationships (Taskar et al., 2002).

4.3.2 Relational Framework

Cognizant of the issues explained in Figure 4.4, with respect to regression, we propose a relational Markov network. The main objective is to estimate the probability of order

interactions (returns) conditional of the level of order imbalance at a given time stamp. In other words, predict returns by using order imbalance and order characteristics information. We represent orders, order imbalance, order characteristics and returns as a relational model. Figure 4.5 depicts the representative relational frame work. As shown in the figure, three templates can represent all the variables. The first template contains order imbalance details that include the respective buy and sell orders and their corresponding stock identification (based on the Location ID which specifies the time interval bucket, the order is from). The second template contains the temporal information relating to order arrival. It contains information on the time of arrival, day of the week, stock identification. The third template contains event that happens due to orders matching algorithm. It contains two boolean attributes namely, interaction and no interaction. Interaction (no interaction) happens if two orders are matched (unmatched).

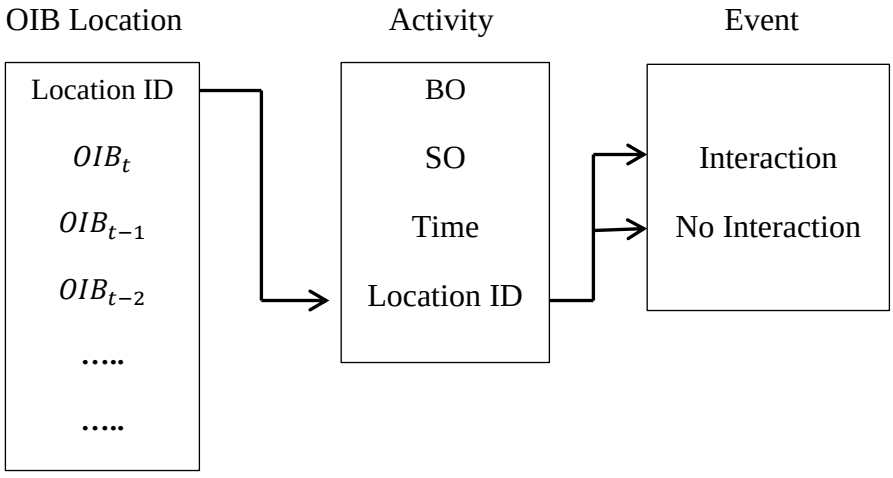


Figure 4.5 Relational Framework

As shown in the Figure 4.5, these three templates are related to each other. The stock identification and corresponding order imbalance are shared between the first two

templates and the activity template is a direct result of the actions that happen in the second temporal template.

RMN defines a conditional distribution p(y|x) over labels y given the observed attributes x. In the thesis, it is the conditional distribution of order interactions y over order characteristics given order imbalance level x. RMN provides a relational joint distribution among all attributes. However, calculating conditional probabilities of each order interaction instance is highly intractable in such complex relational network. We propose a simplification process using Markov Chain adaption, without loss of any generality as described in the next section.

4.3.3 MCMC Method for Missing Data

The Markov Chain Monte Carlo (MCMC) method is used to generate pseudorandom draws from multidimensional and otherwise intractable probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends only on the value of the previous one.

In MCMC simulation, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution, which is the distribution of interest. By repeatedly simulating steps of the chain, it simulates draws from the distribution of interest Schafer (1997). MCMC method has been used by researchers for financial data analysis (Kalimipalli et al., 2004). MCMC has been applied as a method for exploring posterior distributions in Bayesian inference. That is, through MCMC, one can simulate the entire joint posterior distribution of the unknown quantities and obtain simulation-based estimates of posterior parameters that are of interest.

In many incomplete data problems, the observed-data posterior p(θ|Y_{obs}) is intractable and cannot easily be simulated. However, when Y_{obs} is augmented by an estimated/simulated value of the missing data Y_{mis}, the complete-data posterior p(θ|Y_{obs}, Y_{mis}) is much easier to simulate. Assuming that the data are from a multivariate normal distribution, data augmentation can be applied to Bayesian inference with missing data by repeating the following steps.

The imputation I-step: Given an estimated mean vector and covariance matrix, the I-step simulates the missing values for each observation independently. That is, if you denote the variables with missing values for observation i by $Y_{i(mis)}$ and the variables with observed values by $Y_{i(obs)}$, then the I-step draws values for $Y_{i(mis)}$ from a conditional distribution for $Y_{i(mis)}$ given $Y_{i(obs)}$.

The posterior P-step: Given a complete sample, the P-step simulates the posterior population mean vector and covariance matrix. These new estimates are then used in the next I-step. Without prior information about the parameters, a non-informative prior is used. You can also use other informative priors. For example, a prior information about the covariance matrix may be helpful to stabilize the inference about the mean vector for a near singular covariance matrix.

RMN and MCMC in OIB Context: Chapter 3 describes how imputation method improves the estimates of order imbalance that are main predictors of future returns. These estimates are then used for prediction. RMN provides relational framework for orders and their interactions that is more realistic and thus reduces the bias that arises when using regression models. However, as discussed in Section 4.3.2, RMN is a

complex network, as estimation of the predicted values encounter multi dimensionality problem. For instance, as depicted in Figure 4.3 for each order imbalance level template V_1 to V_n, there is an estimated probability or potential that any two orders of two templates interact at the given order imbalance level in the current state. However, due to many templates or order imbalance levels with varied relationships arriving at the equilibrium potential, makes it very complex. This issue is addressed in two steps as explained below.

MCMC method addresses this issue in two steps: In the first step, Markov chain process that follows random walk process simplifies the RMN for estimating conditional probabilities in such a way that the probability of order interaction is conditional only on the current state of order imbalance and independent of any other states in the past. Once such conditional probabilities are arrived, in the second step, Monte Carlo process randomly uses several conditional probabilities as an iterative process until the values converge or gets closer to the equilibrium potential or grand mean.

The next section discusses the algorithms used to achieve the above description of arriving at the equilibrium potential.

4.4 Algorithm

The first step in the algorithmic procedure is to estimate the Order Imbalance level at *t*, a given time interval. Given that, we start with the trade book data that contains only partial order book data, the order imbalance is estimated using the Lee and Ready

(1991)[27] based trade direction algorithm. The underlying order behind the trade is defined as buyer (seller) initiated order if the current price is more (less) than the previous price.

If the current price is lower than the previous price, it is a seller initiated trade (line 1 of Algorithm 1). Then the time interval for the analysis window is calculated (line 2 of Algorithm 1). This time interval is for the selected 5 mins, 10 mins and 30 mins window. Interval in line 2 suggests how the records are grouped based on the interval specified.

Algorithm 1 OIB

Require: $dataset, IntParam$
Ensure: $OIB, Interval$
1: $TD \leftarrow TradeDirection(dataset)$
2: $Interval \leftarrow CalculateInterval(time, IntParam)$
3: $Aggregate \leftarrow Merge(TD, Interval)$
4: $Agg_{Buy} \leftarrow Split(Aggregate)$
5: $Agg_{Sell} \leftarrow Split(Aggregate)$
6: $OIB \leftarrow Agg_{Buy}$ - Agg_{Sell} % Calculate Order Imbalance
7: $OIB \leftarrow MissingValueReplacement(OIB)$

Existing evidence suggests that the return dependency of order imbalance generally will not exceed for more than few hours (Chordia et al., 2005). In addition to that, on average, a typical stock takes few minutes of persistent order imbalance to move stock price away from its equilibrium value. Order imbalance level at minutes or seconds interval may not be large enough to exert any serious pressure on the stock price. Hence, the time intervals considered are 5 minutes, 10 minutes and 30 minutes. The estimated trade direction is obtained by aggregating the values of buyer initiated and

[27] Lee and Ready algorithm may not be necessary in the current ASX context as the historical order book data shows trade direction. However, several stock exchanges do not supply order book data. Therefore, the methodology is for general application.

seller initiated trades for the desired time interval. All the buys and sells are grouped into separate datasets so the buy minus sell is calculated for each interval. This difference of buy minus sell, leads to the estimated OIB for a given time interval. If there are no transactions for a given interval, OIB column will have missing values. These missing values, are then filled with the next interval's OIB value.

The next step is to calculate stock return for a given interval. Algorithm 2 below shows the steps associated with return calculation. For example, if at 11:00 am the stock price is $0.86 and at 11:05 am the price is $0.855, stock return will be the natural logarithmic value of the current price at 11:05 am divided by lag price at 11.00 am (line 1, below).

Algorithm 2 Return

Require: $dataset, OIB$
Ensure: $Final$
1: $Return \leftarrow log(Price/lag(Price))$
2: $LeadReturn \leftarrow CalculateLeadReturn(Return)$
3: $Final \leftarrow Merge(OIB, LeadReturn)$

Line 2 shows that lead return is calculated using the return from line 1. Lead return is the return of the next interval. After the returns and lead returns are calculated for a given time series, OIB calculated in the previous algorithm is merged with returns / lag returns for the same time intervals to generate a final dataset. This final dataset is used to run EM algorithm (Algorithm 3). The final dataset contains the company code, time stamp for the given time interval, OIB in terms of number, OIB in terms of volume, OIB in terms of value and their corresponding returns. From this dataset, a percentage of records are randomly deleted to create a dataset of *MissingOIB*. These records are then imputed using EM Impute. EM uses MCMC procedure to draw the pseudo random

draws from the distribution and estimates conditional probabilities by iterating until the values converge to a grand mean.

Algorithm 3 EM Impute

Require: $Final, OIB$
Ensure: $MissingOIB, ImputeOIB$
1: $MissingOIB \leftarrow rand(pPercent, OIB)$
2: $ImputeOIB \leftarrow EMImpute(Return, MissingOIB)$

The MI (multiple imputation) procedure in SAS provides the method to create imputed datasets. The EM statement in the proc MI, uses the EM algorithm to compute the maximum likelihood estimate (MLE) of the data with missing values, assuming a multivariate normal distribution for the data.

The code below is used to impute missing data by EM; the second line of the code below uses "em" on variables: return, miss_OIB_Volume. Return and Missing OIB are the variables (VAR in the code) considered for analysis.

PROC MI data= < dataset used > ;

EM < options > ;

VAR variables ;

Once the missing values are imputed, the outputs of Algorithm 2 and Algorithm 3 are used to calculate the regression estimates in Algorithm 4. The dependant variables of the regression are return and OIB for complete data; return and MissingOIB for randomly missing data; return and ImputeOIB for EMImputed data.

Algorithm 4 Regressions

Require: $Final, MissingOIB, ImputeOIB$
Ensure: $LeadReturnF, LeadReturnM\ LeadReturnI$
1: $LeadReturnF \leftarrow Return, OIB/adjrsq$ % Full Regression
2: $LeadReturnM \leftarrow Return, MissingOIB/adjrsq$ % Missing Regression
3: $LeadReturnI \leftarrow Return, ImputeOIB/adjrsq$ % Imputed Regression

Using these three estimates, regressions[28] are run to compare the estimates and to validate the value addition of EM imputation procedure. Before we implement the algorithms, it is important to understand the relationship between orders, order imbalance and returns. In the next section, we develop a theoretical model that establishes the relationships and later develop theoretical and empirical estimation procedures.

4.5 A Relational Probabilistic Model

The main objective of the probabilistic model is to establish the distributional properties of order characteristics and their interactions. Once we define the theoretical distribution of order characteristics, order interactions and also their joint probability distribution function, we can estimate the expected value or the mean of the distribution. This expected value serves as a starting point for the empirical analysis.

We follow the notations as shown in the Table 4.1 to understand the discussion on the distributions of order characteristics, their interactions and the join probability distributions. We assume that each order *o* belongs to one of *k* order imbalance levels. Where, an order imbalance level represents a particular magnitude of order imbalance that causes a particular level of trade direction or order interaction that results in potential change in stock price. For instance, on average, for every 0.1% of positive order imbalance, OIB (buy orders more than sell orders) at time "*t*", the trade direction at "*t+1*" would be, on average, positive 10% appreciation in stock price. Let us assume

[28] The correlation issues relating to the regression model will still remain when we use our model when comparing with the existing regression model (Chordia et al., 2005). These regression related estimation issues are addressed using GMM and Fama Macbeth procedures that are discussed in the experimental setup of Chapter 5.

there are *C* distributions of orders each belonging to a specific template of *k* order imbalance levels. This can be represented as $o.C \in \{1,....,k\}$. It has to be noted that order book is not visible, so one cannot observe the *level* of OIB.

Table 4.1 Table of Notations

o	Each order
k	Order imbalance level
C	Total distributions of k levels
$o.C$	Order in the distribution
$o.C \in \{1,....,k\}$	Order in the distribution belongs to a particular OIB level *1 to k*
o_i	A given instance *i* of an order
$o_i.C$	A given instance *i* of an order in the order distribution
E	Expressions of order
m	Number of expressions
j	An attribute belonging to *m*
$o.E = \{o.E_1,....,o.E_m\}$	Each order has a number of expressions from *1 to m*
$o.E_j$	Specific attribute of the order; Eg: attribute - price level at which the order is placed and belongs to the set $o.E$
θ [29]	Assigned probability value
θp	Probability that a given OIB level in $o.C$ is equal to p
$V = \{V1, \ldots, Vn\}$	Set of *n* discrete random variables
V_i	Interacting attribute *i* within the random variable V_n
V_j	Interacting attribute *j* within the random variable V_n
$\phi\, i$	Interaction potential of attribute *i*
$\phi\, j$	Interaction potential of attribute *j*
[*Vi*—*Vj*]	An edge where attributes *i* and *j* within two random variables have a possibility to interact
$\phi\, i, j$	Compatibility potential of two attributes *i* and *j* within two *V*s
oi, oj	Orders corresponding to Vi and Vj
θ [30]	Mean estimate of OIB
Θ	Posterior distribution which is multinomial θ

[29] In Section 4.5.1.

[30] In Section 4.6.

In other words, one cannot observe the organizational structure of the traders that causes a particular level of OIB. Therefore one can say that an OIB level, for a given instance of order o_i from the order distribution *C* denoted by $o_i.C$, is latent (or hidden). This value needs to be estimated. In order to estimate this value, first we model the distribution properties of order characteristics and order interactions separately (the detailed procedure of estimating order imbalance level is explained in Chapter 3). Later, we combine these two for estimating the joint distribution of order characteristics and their interactions to provide a single unified model.

4.5.1 Modelling Order Characteristics

In this section we define the distributional characteristics of orders. As discussed in the Chapter 2, we use Gaussian Processes (hereafter GP) as the base methodology due to its superiority among machine learning applications for analysing financial time series data and its ability to capture missing data. As per GP methodology, we assume that distribution of orders is a multinomial distribution with several order imbalance level based distributions. Let $o_i.C$ be the order instance, for a given level of order imbalance from the order distribution *C*, where the order belongs. Each order instance has a number of expressions. This number is denoted by *m* and the expressions are denoted by *E* and are same for all orders. So, each instance of *o* has *m* continuous-valued different expressions.

Let j be one of the expressions (say, price level) and belongs to $o.E = \{o.E_1, ..., o.E_m\}$ where $o.E_j$ represents price level at which the order is placed and belongs to the set $o.E$.

The naive Bayes model defines probability distribution of order expressions $o.E$ that belong to $o.C$ is as follows:

$$P(o.C, o.E1, \ldots, o.Em) = P(o.C)\prod_{j=1}^{m} P(o.Ej \mid o.C) \qquad (4.2)$$

Equation 4.2 represents the probability distribution function of order expressions that belong to (*o.C*) for a level of order imbalance. The random variable *o.C* is distributed as a multinomial distribution,[31] parameterized by the vector of order imbalance levels $\theta\mathbf{C} = \{\theta\mathbf{1}, \ldots, \theta\mathbf{k}\}$, where θ is the assigned probability for the k possible finite outcomes (k possible levels of order imbalance) and $P(o.C = p) = \theta p$; thus, each $0 \leq \theta p \leq 1$ and $\sum_{p=1}^{k} \theta p = 1$ (total probability equals 1). We model each *conditional probability distribution (CPD) P (o.E j | o.C = p)* using the Gaussian distribution $N(\mu_{pj}, \sigma_{pj}^2)$. [32]

4.5.2 Modelling Order Interactions

To be consistent with the RMNs framework discussed in Section 4.2.2, we use the framework of *Markov Networks* for modelling the distribution of order interactions. Let $V = \{V1, \ldots, Vn\}$ be a set of discrete random variables of order imbalance levels from 1

[31] Multinomial distribution accommodates the possibility of capturing missing data by allowing several normal distributions under the same multinomial distribution with different expected values (order imbalance levels).

[32] We use Gaussian distribution to be in line with the RMNs framework.

to n. A binary Markov network over *V* defines a joint probability distribution *P(V)* as follows. The network is defined via an undirected graph whose nodes correspond to variables in *V* and whose edges *E* represent direct probabilistic interaction between those variables. Variables *Vi, Vj* are associated with potentials ϕ *i (Vi)* and ϕ *j (Vj),*where ϕ *i,,* ϕ *j* are the interaction potentials. Each edge between variables *Vi* and *Vj* represented as [*Vi—Vj*] is associated with a non-negative *compatibility potential* ϕ *i, j (Vi , Vj).* Note that ϕ *i (Vi),* can be derived from Equation 4.2.

The joint distribution is then defined as

$$P(V1, \ldots, Vn) = \frac{1}{Z} \prod_{i=1}^{n} \phi i(Vi) \prod_{[Vi - Vj] \in \varepsilon} \phi i, j(Vi - Vj) \qquad (4.3)$$

where *Z* is a normalizing constant defined so as to make the distribution sum to 1. Intuitively, ϕ *i (Vi)* encodes how likely the different values of *Vi* are, ignoring interactions between the variables. For an assignment *oi*, *oj* to *Vi* , *Vj* (where, *oi, oj* are the orders in given levels of order imbalances *Vi , Vj*) , the value ϕ *i, j (oi, oj)* specifies how 'compatible' the assignment *oi, oj* is: the higher the value, the more likely this pair of values is to appear together.

In this setup, the variables *V* are OIB levels *o*1.*C, . . . , o.C* of orders in O, and edges correspond to observed buy-sell orders interactions. Intuitively, an edge between *oi* and *oj* captures the basic intuition that, if *oi* and *oj* interact, they are more likely to belong to a certain level of OIB or *V*. Thus, we define the compatibility potential ϕ *i, j (oi .C = p, oj .C = q)* such that the compatibility value for $p = q$ is greater than the value for $p \neq q$. Since we do not assume any patterns over the distribution of interactions, we

set all entries in which $p = q$ to the same value. Similarly, all entries in which $p \neq q$, are set to a same value. This classifies interactions and non-interactions data. Due to the normalization of the distribution, what matters is only the relative magnitude of these two values. Thus, we can parameterize the interaction model using a single parameter, α, such that for all

$$\phi_{i,j}(o_i.C = p,\ o_j.C = q) = \begin{cases} \propto & p = q \\ 1 & otherwise \end{cases} \tag{4.4}$$

α represents strength of preference towards assigning interacting orders that belong to the same OIB level and takes values greater than 1. We can observe the interactions of orders from the trade book and we can also measure the strength of coherence through correlation between orders that are assigned to the same OIB level. The larger the value of α implies more orders interact for a given level of order imbalance. Given a database of trades that define the set of edges E, the parameterizations for ϕ_i and $\phi_{i,j}$, Markov network defines a joint distribution $P(o_1.C, \ldots, o_n.C)$ over assignments of different levels of order imbalances. The value for α could be a range of assigned values based on a number of order interactions and the coherence of order expression profiles in a given level of order imbalance. For instance, we can set different levels of α based on number of trades that occurred at a particular level of order imbalance. Likewise, we can also use different correlation levels for setting values for α. Now, by combining distributions $P(o_i.C)$ (Equation 4.3) as the potential $\phi_i(oi\ .C)$ in the interaction model) and $\phi_{i,j}(o_i\ .C, o_j\ .C)$, the combination can be considered as an RMN. RMN combines two distributions of order imbalance levels.

The result of combined model defines a joint distribution over the entire set of random variables, as follows:

$$P(O.C, O.E|\varepsilon) = \frac{1}{Z}\left(\prod_{i=1}^{n} P(oi.C)\prod_{j=1}^{m} P(oi.Ej \mid oi.C)\right).\left(\prod_{[Vi-Vj]\in\varepsilon} \phi i, j(Vi - Vj)\right) \quad (4.5)$$

where Z is a normalizing constant that ensures that P sums to 1, and E represents all binary interactions that exist between orders in the data. $\mathcal{E}$ represents set of edges and all $[Vi—Vj\ ] \in\ \varepsilon$. In summary, Equations 4.2 and 4.5 provide distributional properties of order characteristics and their interactions. These distributional properties can be used as the starting point for addressing missing data problem. The expected values based on these distributions are used as the initial estimates in our following proposed machine learning methodology. Equation 4.2 gives the distribution of order characteristics. Equation 4.4, classifies interactions and non-interactions data. After the interactions are known, the distributions of these interactions when combined with the distributions of order characteristics (Equation 4.3), form the joint distribution. Relational Markov Network Model is defined for estimation of the joint distribution function of orders, order characteristics and their interactions. The grand mean of the distributions is calculated using EM algorithm in the empirical setting. This is explained in the next section, Section 4.6 through E-step and M-step. This estimate of the grand mean gives the imputed data estimates of Algorithm 3 of Section 4.4.

4.6 Learning through Expectation Maximization Algorithm

The probability distribution functions (Equations 4.2, 4.3 and 4.5) in the previous sub-sections are aimed at calculating the expected values with an assumption that data are

complete. However, the available data in trade book have missing orders from the order book. Hence, in order to have true estimates of the parameters we have to re-estimate by including the missing values. One common solution for this problem is to calculate the maximum likelihood estimate by changing the mean or expected values. However, it is not possible to manually engage in a trial and error estimation method.

Machine learning research advancements aid in obtaining optimal solutions. One of the most widely used machine learning method for obtaining such maximum likelihood estimator is Expected Maximization Algorithm (EMA).[33] EMA is frequently used for data clustering in machine learning and computer visualisation applications (Gonzales and Woods, 2002). EM is also used to study missing data in finance in the analysis of time series data (Hamilton, J.D., 1990, Warga, A., 1992 and Zhu, H., 2006).

This method was originally proposed by Dempster et al. (1977) and later a formal statistical framework was created by Little and Rubin (1987). The authors used multinomial distribution framework (that suits our description of order imbalance levels) to address how missing data can be included in incremental steps for ensuring better maximum likelihood estimates at each iteration. In other words, Little and Rubin (1987) developed EMA as a generalization of maximum likelihood estimation to the incomplete data sets. This statistical framework ensures generalizability to any missing data problem that is missing not at random. In the words of Schafer (1999), "If we knew the missing values, then estimating model parameters is straight forward. Similarly, if we

[33] For example, Ghahramani and Jordon (1997), Mclachlan and Krishnan (1997), and Gales and Young (2008).

knew the parameters of the model, then it would be possible to obtain relatively less biased predictions for the missing values".

The EMA process explained below is implemented in the Algorithm 3 (in Section 4.4) by the MI procedure of SAS which uses the EM method. The flow of EMA process can be decomposed into three major steps. Figure 4.6 depicts these three steps.

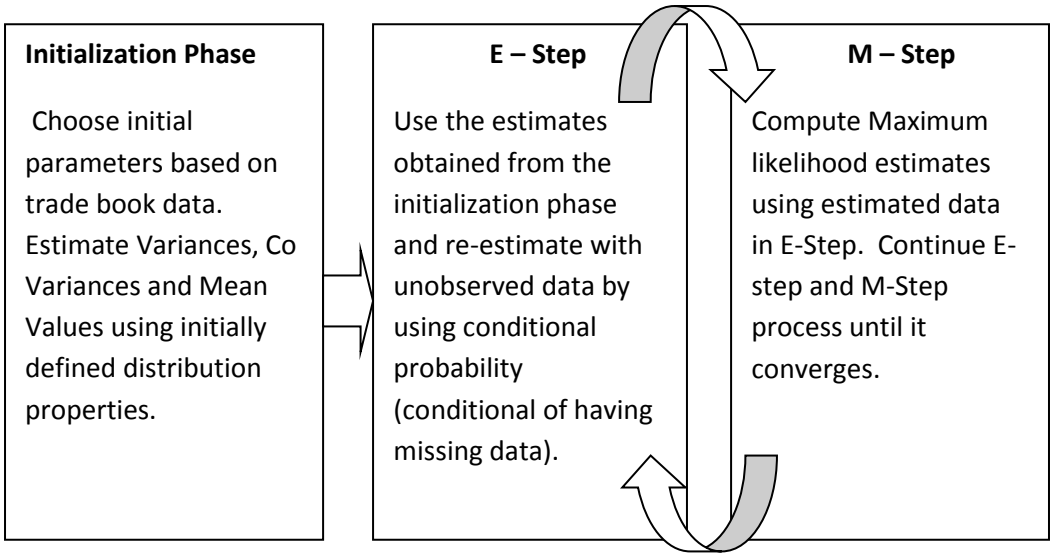


Figure 4.6 EMA Estimation Procedure

In Step 1, the initialization step, we need to assume that the data available to us are complete and estimate the mean, variance and co-variance as the initial parameters. In our case, we use Equation (4.5) based estimates of means, variances and co-variances as initial parameter estimates, using the available trade book data. For instance, trade book gives information on orders (that have interactions) and traded order imbalance. Using this information, we can calculate the determinants of future returns. Given that order imbalance based strategies need to be implemented in minutes or seconds intervals, inaccuracy of predictions can be very costly. In the EMA process, these initial estimates are used as the input values into the step 2 i.e., E-Step. The E-Step computes

the distribution over hidden variables using the initial estimates as the parameters. In a simple way, it can be thought as running a regression $Y\hat{} = a + bx$ [34] with the data we have, then use x to estimate $Y\hat{}$ wherever it is missing.

This computation is based on conditional probability of including missing data in the estimation process. Based on the new estimate of E-Step, we calculate new maximum likelihood function in step 3 i.e., M-Step. The sufficient statistics (mean, variance and covariance) obtained in M-Step are used as the new parameter values in the iterative E-Step. This process between E-Step and M-Step iterates until convergence or until arriving at the best estimates (closest to the grand mean). It is to be noted that, new set of information is added to run the M-Step and hence some new missing data get captured for estimation at every step. The technical description of the intuition behind EMA process is discussed below.

EMA tries to attempt finding the parameter $\hat{\theta}$ that maximises the log probability $\log p(x;\theta)$ of the observed data. Unlike the case of complete datasets, in the case of incomplete datasets, the objective function $\log p(x;z;\theta)$ includes z to represent hidden variables. Given this complexity, there cannot be a global maxima that often be computed in a closed form solution (Little and Rubin, 1987). EMA addresses this issue by reducing $\log p(x;\theta)$ into a sequence of simpler optimization sub problems that can have unique global maxima and that can often be computed in closed form. These sub-problems are chosen in a way that guarantees their corresponding solutions $\hat{\theta}^1, \hat{\theta}^2$ …., converge to a local optimum of $\log p(x;\theta)$.

[34] $Y\hat{}$ represents return, R_t and x is a vector of two determinants: OIB_{t-1}, R_{t-1}.

EMA alternates this in two steps. In step one, the E-Step, it chooses a function g_t that lower bounds $\log p(x;\theta)$ everywhere for which $g(\hat{\theta}^{(t)}) = \log p(x;\hat{\theta}^{(t)})$. In M-step, the EMA moves to a new parameter set the $\hat{\theta}^{(t+1)}$ that maximises g_t. As the value of lower bound g_t matches the objective function at $\hat{\theta}^{(t)}$, it follows that the $\log P(x;\hat{\theta}^{(t)}) = g(\hat{\theta}^{(t)}) \leq g(\hat{\theta}^{(t+1)}) = \log p(x;\hat{\theta}^{(t+1)})$.

This implies that, for each iteration, there will be a monotonic increase in the value of the objective function. We adopt EMA to map hidden orders with an assumption that order expressions and their corresponding interaction can be traced back to a specific level of order imbalance. In our case, $o_i.C$ are hidden, and are learnt from the data at the same time as the parameters. Let O be a set of orders, and assume that our observable trade book dataset contains: for each order o_i, an order expression profile (for instance, price and order size); and a set of order interactions E between pairs of orders (o_i, o_j). The main objective is to learn parameters Θ, which consists of: the multinomial θ over a specific order imbalance level, and means and standard deviations of each of the *k* Gaussian distributions (order imbalance levels) associated with each of the ($k \cdot m$) conditional probability distributions (CPD) *P(o.Ei* | o.*C* = *p)*. The E-Step computes the distribution over the hidden variables given the observed data by using the current estimate of the parameters. In other words, we compute $P(O.C \mid D, \Theta^{t-1})$. To compute this distribution, we must run inferences over Equation (4.4). Using posterior distribution, the M-step re-estimates the model parameters using expected sufficient

statistics (such as, mean, variance and covariance), where the expectation is taken relative to this posterior. Thus, parameters are estimated.

It is to be noted that, multiple imputation (MI) procedure of SAS that uses EM algorithm, does not attempt to estimate each missing value through simulated values, but rather to represent a random sample of the missing values. Instead of filling in a single value of the estimated parameters for each missing value of OIB, multiple imputation procedure replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute (Rubin 1976, 1987).

The multiple imputed datasets are then analyzed by using standard procedures for complete data and combining the results from these analyses. No matter which complete-data analysis is used, the process of combining results from different data sets is essentially the same. This process results in valid statistical inferences that properly reflect the uncertainty due to missing values; for example, valid confidence intervals for parameters.

Once the estimates are ready, we run regression on complete dataset, missing dataset (dataset after random deletion) and imputed dataset and compare the regression results to validate the value addition of the proposed methodology. These regressions are implemented through Algorithm 4, discussed in Section 4.4. Thus, we estimate the parameters and the distribution for a specific order imbalance level and address the research question of mapping the hidden orders for drawing more accurate inferences on trade direction.

4.7 Summary

The objective of this chapter is to propose prediction method for estimating the extent to which order imbalance can predict future returns. Cognizant of the problems associated with missing data and the relational complexities associated with order imbalance and returns, this chapter addresses these issues in two steps. First, unlike standard regression methodology, this chapter proposes a Relational Markov Network Model for estimation of the joint distribution function of orders, order characteristics and their interactions. Second, the Expectation Maximization Algorithm is proposed to address the missing data problem during the joint estimation procedure.

The proposed novel methodology overcomes the estimation problem, especially in the context of missing order book data. Regression models do not identify the relationship between order imbalance and returns. Orders and their imbalance at the given point of time influence the price movements (and the corresponding returns) and such interdependencies, if not jointly estimated, lead to biased estimates. RMN framework allows us to overcome this problem by establishing the relational network between variables and further establishing joint probabilistic functions. EM algorithm then maximizes the likelihoods of the estimates by learning information related to the missing orders through posterior distribution functions and thus improving the accuracy of the estimates.

In summary, the proposed machine learning based methodological framework that can be coined as Algorithm for Imputed Complete Order Book (AICOB) ensures more accurate prediction than the exiting regression models used in finance literature. With clearer representation of the relation between orders and returns and addressing the

problem of cross correlations that arise in the standard regression models the predictive ability can be enhanced.

This chapter leads us to undertake empirical tests to provide evidence for validating the marginal benefit (over established empirical methods in finance literature) of using the newly proposed methodological framework for order imbalance based prediction.

This page is intentionally left blank.

Chapter 5

EFFICACY OF THE ALGORITHM FOR IMPUTED COMPLETE ORDER BOOK (AICOB)

5.1 Introduction

Evaluating the efficacy of our proposed Algorithm for Imputed Complete Order Book (AICOB) is our last and critical step in the thesis. One of the main objectives of the thesis is to find out whether missing data problem associated with trade book is solved through AICOB procedure. Second, to understand whether AICOB based lagged Order Imbalance (OIB) estimates predict stock returns better than the existing methods of estimating OIB (Chordia and Subrahmanyam, 2004) that use trade book data, which suffer from missing data problem.

The conventional efficacy measure of an algorithm in Information Technology area revolves around understanding the time complexity (the amount of time taken by the algorithm to complete the task; it is commonly expressed in computer science using big O notation) and resource consumption (how much load is levied on the systems). In our case, efficiency of AICOB is measured not in terms of time complexity, instead, in terms of duration of prediction power. The longer the algorithm predicts future return for, the higher would be the potential to earn profits, through trading. Hence, efficiency is defined as the time duration that allows traders to trade and obtain potential economic profits from the algorithm.

Resource consumption is not a major issue with the advancement of technological innovations. Most of the trading firms use super computers and engage in regular algorithmic based trading (Hendershott et al., 2011). Hence, duration of prediction becomes critical while competing with other algorithmic traders. Several researchers use indices like Rand Index as benchmarks for evaluating the accuracy (Aghabozorgi, Shirkhorshidi and Wah, 2015). Given that we have access to the complete order book data, which is the true benchmark, we use this complete order book data based predictions as our benchmark for evaluation. Under this dimension, we test the efficacy of the algorithm proposed in Chapter 4, Section 4.4 for its accuracy in predicting stock returns. In Section 4.4, the 4 steps are shown as 4 algorithms. These four are combined together and termed as AICOB. This analysis meets the second objective of the thesis. We test how close AICOB results are with respect to the complete order book results. Finally, following finance literature, we conduct sensitivity analysis by varying firm size based parameters to check whether the algorithm is adaptable to these varying dimensions.

In this chapter, we report our empirical results relating to future returns prediction through OIB data. Before we embark on the implementation of the empirical methodology discussed in Chapter 4, we design the evaluation strategy in Section 5.2 that will ensure that the research objectives are met. Evaluation strategy is presented first by decomposing into three major dimensions, namely, efficiency, accuracy and adaptability. These three dimensions ensure a thorough investigation on the efficacy of the algorithm for practical use in the real-world trading environment. Efficiency focuses

on the economic aspects of the methodology. Under accuracy dimension, we focus on how close the estimates are to the actual complete order book data (historical).

This is a major contribution of our study, as the current literature does not compare the results with the actual complete order book. The third dimension, adaptability, based on sensitivity analysis focuses on the applicability of the proposed methodology to the real world. The empirical results are presented in three sections, covering each dimension of the evaluation strategy. The introduction is followed by a description of the evaluation strategy. This is followed by the discussion in Section 5.3 on the measurability of the goodness of the prediction under each of the three evaluation methods. Here, we define our expectations on predictability under each dimension.

Next, in Section 5.4, we introduce the data that are being used for our empirical tests. The huge data size and the corresponding complexities are highlighted to comprehend and visualize the need for machine learning methods in financial applications. We then discuss the implementation steps of AICOB in detail in Section 5.5. This is important to track the various procedures involved in obtaining the final results. In Section 5.6, the data distribution statistics of size based groups are discussed. Section 5.7 contains the experiments conducted, using the Australian stock market data to address the research questions and the corresponding results are reported under the evaluation dimensions in Section 5.8. The results are further classified into sub sections under each evaluation dimension for better tractability. Finally, we interpret the results in Section 5.9 and conclude the chapter in Section 5.10.

The existing research papers (Chordia et al., 2005) generally use pooled cross section data with no corrections for cross sectional and time series correlations. One of the important contributions of the thesis is that all pooled regression results of AICOB follow *Fama MacBeth* (Fama and MacBeth, 1973) and Generalized Methods of Moments (GMM) procedure and are discussed in this chapter. These methods control the cross sectional and time series correlations between the observations and across the pooled stocks. Thus, the objectives of the thesis are achieved by answering the research question on whether machine learning-based algorithm can address missing data problem in financial time series data and whether they help in predicting the future returns better than that of conventional methods used in the current finance literature. On the empirical front, the main contribution of this chapter is to provide comparison with order book data that support our claims that AICOB's results are closer to the future returns predictions of order book.

5.2 Evaluation Strategy

Evaluation of algorithm, as discussed in the previous section, is based on measurable objective criteria, that covers three important dimensions, namely, economic benefit or how long AICOB based lagged OIB predicts returns (*efficiency*), quality of the algorithm or how close the estimates based on AIOCB data are compared to the complete order book based estimates (*accuracy*) and how consistent the predictions are to variations in firm size (*adaptability*). These three dimensions are used in evaluating the empirical methodology proposed in Chapter 4.

In the context of the thesis, we evaluate "can the return prediction improve by using the proposed methodology compared to that of the existing methodology used in the literature?" This process is carried out by comparing the trade book results (that is incomplete) with the AICOB results (based on the Chapters 3 and 4) and then with the complete order book results. Figure 5.1 illustrates the evaluation process.

As shown in the Figure 5.1, A1 represents complete order book data that includes all orders with no missing data. B1 represents full trade book data and B2 represents imputed data (AICOB). The B1 with B2 comparison represents an evaluation of the AICOB (that uses B2) with standard regression methodology (that is used on B1) to evaluate the superiority of B2 based results over B1 based results. Both B1 and B2 are compared to A1 (complete order book). These comparisons together form the basis for evaluation in terms of efficiency, accuracy and adaptability.

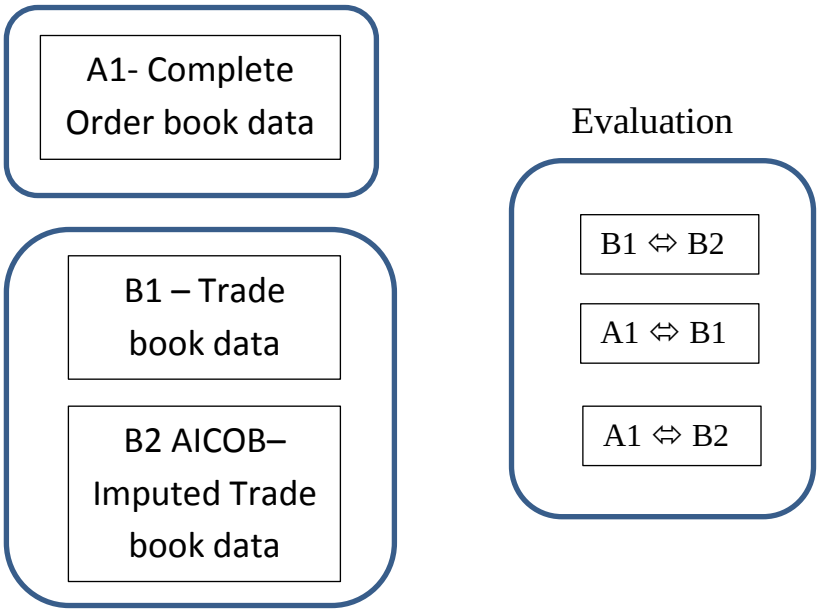


Figure 5.1 Comparisons of Order Book and Trade Book for Evaluation

5.2.1 Accuracy

The research question addressed for accuracy is, *how good are the AICOB based estimates compared to the complete order book based estimates?*

Accuracy is measured by comparing the estimates of imputed order imbalance coefficients calculated from the trade book data with estimates of actual order imbalance data from the complete order book. The accuracy of the AICOB is evaluated by comparing the closeness[35] of the imputed data estimates (B2) with the complete order book estimates (A1). The closer the imputed order imbalance data are to the actual order imbalance data, the higher is the accuracy of the AICOB.

5.2.2 Efficiency

The research question addressed for Efficiency is, *how long the predictions based on AICOB estimates, last for?*

Efficiency is measured here by calculating the estimates for different time windows. Following the standard conventions of intraday return prediction (Chordia et al., 2005), the time windows used are 5 minutes, 10 minutes and 30 minutes intervals. The longer the predictive ability of the AICOB based data, the better is the efficiency. The efficiency of the AICOB is evaluated by comparing whether prediction based on B2 lasts longer than B1, where B2 is closer to the prediction based on A1 (compared to B1).

[35] Closeness is defined by comparing the estimates of coefficients and t-values.

5.2.3 Adaptability

The research question addressed for Adaptability is, *can the AICOB handle variations in stock size?*

Finance literature has both theoretical and empirical evidence that OIB predicts future returns for large firms more effectively compared to small firms (Chordia and Subrahmanyam, 2004). This is due to longer order history dependency for large firms, where competitive traders place many orders that are closer, in terms of price and quantity, to each other. Hence, we test whether AICOB is adaptable to firms of varying size. Adaptability is measured by estimating the predictive ability of the AICOB for stocks of different market capitalization.

Evaluation is also carried out by comparing whether B2 based results are more robust to variations in stock size compared to B1. This analysis helps to understand whether the result based on AICOB vary with the conventional results that are carried out by using B1 data. Several inferences based on existing studies are drawn based on B1 data. Hence, B2 data that imputes (with AICOB estimates) the missing data of trade book can uncover the variations of duration based prediction that can be attributed to missing data problem.

5.2.4 Summary of the Evaluation Strategy

Table 5.1 summarizes the discussion on the evaluation strategy. The table has four columns to report the criteria of evaluation, the research question that is being addressed through evaluation, the procedure of evaluation and expected outcome of evaluation,

respectively. The evaluation, as discussed earlier, is based on accuracy, efficiency and adaptability of the AICOB.

Table 5.1 Overview of the Evaluation Strategy

Evaluation Criteria	Research Question Addressed	Procedure	Outcome
Accuracy of the AICOB	Does the AICOB accurately predict estimates close to the estimates generated from the actual hidden data in the order book?	Estimate predictive ability based on regression procedure using both hidden data with AICOB and actual complete data.	Comparisons of the results indicate the accuracy of the AICOB in terms of the prediction correctness.
Efficiency of the AICOB	How long the predictions last for engaging in trading effectively?	Estimate predictive ability of lagged OIB for different time windows. For instance, for 5 minutes, 10 minutes and 30 minutes intervals.	The longer the duration of the prediction, the better is the efficiency of the AICOB.
Adaptability of the AICOB	Does the AIOCB handle variations in terms of variation in stock size?	Estimate predictive ability of lagged OIB, in terms of both accuracy and efficiency by changing stock size.	Larger stocks with deep order books (high trading activity) should be more predictable as the issue of missing data for trade book is most prevalent for these stocks (as discussed in Section 3.6).

For each evaluation strategy, we use the complete order book data as the benchmark. Under the accuracy dimension, we expect that the AICOB based estimates are close to the complete order book based estimates. As shown in Table 5.1, the data are evaluated to see if imputed data captures hidden data from the order book that is missing in the

trade book data. As part of the evaluation strategy, we next proceed to the goodness of prediction strategy to emphasize what is expected as the right or good prediction within the context of finance theory.

5.3 Goodness of the prediction strategy

Our goodness of the prediction is guided by the finance literature (Chordia and Subrahmanyam, 2004) as the foundation to define what is expected to be a good prediction. The following paragraphs provide detailed description and justification of the parameters chosen.

5.3.1 Firm Size and Order Imbalance

Firm size generally dictates the trading activity of the concerned stock. Larger firms have more orders and higher competition among traders. In the case of smaller firms, due to lower number of traders and lower competition, orders placed by traders will not be that closely related to each other. Chordia and Subrahmanyam (2004) use firm based quintiles to understand the relationship between order imbalance and stock returns. They find that stock prices of largest firm size quintile react most quickly to order imbalance. This implies that large stocks, being highly liquid exhibit higher history-dependent order imbalance effect. On the other hand, for smaller stocks, this effect is opposite. The small stocks, with less trading, exhibit lower history-dependent order imbalance effect.

5.3.2 Adaptability of the AICOB

The adaptability of the AICOB is tested by dividing stocks into three groups based on their size, as small stocks, medium stocks and large stocks, respectively. Stock market

capitalization[36] is used as the measure of firm size. As discussed earlier, small firms should exhibit lower predictive ability due to less history-dependent order imbalance. On the contrary to the conventional wisdom, large stocks, in the missing data context, can be predicted better as they have less missing values compared to small stocks. Hence, the prediction will be lower for small stocks. Likewise, large size firms should exhibit higher predictive power due to large pile of history-dependent order imbalance (Chordia and Subrahmanyam, 2004). However, it is important to note that, imputed values, using the AICOB, are mainly inferred from the missing data that is not visible in trade book.

Given that small size firms, due to their thin trading, carry only a few high information trades, it is less likely that they will have history-dependent order imbalance. Hence, even with few missing trades, inference significance can drop significantly and thereby imputed order imbalance values may not act as robust predictors. This implies that, for small stocks, the imputation procedure may not exhibit higher predictive power as predicted in the finance theory.

5.4 Australian data

We have taken the Australian stock market data for our experiments and evaluation for two important reasons. First, the Australian stock market is an order driven market, to which the algorithm can be applied. Second, Monash University officially subscribes[37]

[36] Market capitalization here refers to the market value of a company.

[37] Data provider: SIRCA – Securities Industry Research Center of Asia Pacific.

to the Australian dataset and hence can provide better data support system. The data variable definitions of both order and trade book are reported in Table 5.2.

Table 5.2 Australian Data Definitions

Data Variable	Definition
Instrument	Type of security instrument being traded in the stock exchange. For instance, equity, Index contract, debt, futures or option
Date	Transaction Date
Time	Transaction timestamp up to 100 milliseconds
Record Type	Whether the transaction is a trade or an order
Price	Stock traded/ ordered price in $ AUD
Volume	Number of shares traded or ordered
Value	Product of price and volume
Trans ID	Unique transaction identification number
Bid ID	Unique bid identification number
Ask ID	Unique ask identification number
Bid/Ask	Whether the order is a bid or an ask

The data span over 12 months (year 2012) that occupies around 1 Terabyte space and contains around 300 million rows. Given the sheer size of the dataset and keeping computing resources in mind, it is decided that the analysis will be performed for two

months. We used January and February months of the year 2012[38] for the whole analysis. The final sample of two months data occupies around 18 Giga bytes and contains around 60 million rows. The analysis is based on both trade data and order data. Trade data contain around 7 million rows in our final sample. This is about 12% of the total data. This highlights the fact that about more than 80% of the data are missing in trade book.

5.5 Implementation Steps

This section shows the process of our implementation procedure. The implementation process involves four major steps. The first task is cleaning up and arranging the huge data into a format conducive to computing. The second task is to set up the trade book based OIB estimates using existing financial research based methods (Lee and Ready, 1991). The third step is the actual implementation of AICOB that is described in Chapter 4. The final step is to check the process of regression estimates' robustness through *Fama MacBeth, GMM* procedure. Figure 5.2 provides the visual description of the implementation process and the following paragraphs describe the steps with reference to Figure 5.2.

Step 1 is the initial step that involves data pre-processing and database creation. As shown in Table 5.2, the data contain several types of instruments and several types of records. We selected only equity securities and trade data for our analysis. Later, we delete firms that are highly illiquid with very few trades in the sample period.

[38] Started my research by using ASX data in the year 2012. However, for intraday analysis, due to millions of high frequency observations, year or month related effects should not be of any significance.

Step 2 involves implementation of the Lee and Ready (1991) algorithm for obtaining whether a trade is initiated by buyer or seller. This algorithm is required for calculating order imbalance at a given trade frequency. Next, we create a variable list for running regression. Our main objective is to test whether order imbalance, after controlling for past returns, can predict future returns (Chordia et al., 2005). Hence, we need time series data of contemporaneous returns, lagged returns and lagged order imbalance at different time frequencies for each firm. We created mainly 5 minutes, 10 minutes and 30 minutes frequencies of time series data for our analysis (Chordia et al., 2005).

Step 3 involves in using the frequency based time series data for running the AICOB procedures in SAS. We use the Monte Carlo Markov Chain (MCMC) procedure for obtaining samples of unknown variables from the joint posterior distribution of our estimation equation. MCMC is a popular procedure in Bayesian framework where densities of variables are estimated while dealing with huge datasets that often have multi-dimensionality problem during estimation. Later, we run the Expectation Maximization (EM) procedure based on multiple imputations method. Data are assumed to have missing at random setup (MAR, discussed in Section 3.3). Lagged order imbalance observations are randomly deleted and the Expectations Maximization (EM) procedure is run to estimate order imbalance based on the observed lagged return data. We use full data based estimation for benchmark purposes. This procedure provides firm level estimates (one estimate for each firm in the sample) of Lagged OIB to understand the predictive power of such imputed estimates.

Given that all the estimates are at firm level, for obtaining robust aggregate estimates, we use the Fama MacBeth (1973) procedure. We need an aggregate estimation for general inference of EM procedure. This procedure provides single estimate based on several multiple firms over time by using two steps. First, firm level time series beta estimates of predictors are obtained and these estimates are used as inputs in the second step. Next, we run cross-regression by pooling individual time series estimates. This procedure is popularly called as the Fama MacBeth (1973) estimation procedure. Given the cross sectional regression property of the second step, the standard errors of the traditional Fama MacBeth (1973) are corrected only for the cross-sectional correlation. For correcting standard errors of time series regressions, we need to run Generalized Methods of Moments (GMM). Hence, we use GMM estimation to make more meaningful inferences on the pooled-cross sectional time series data.

Step 4 is the final step of our analysis. Here, we run the robustness check by size based segregation of the data to observe the consistency of the predictive power through iterative process.

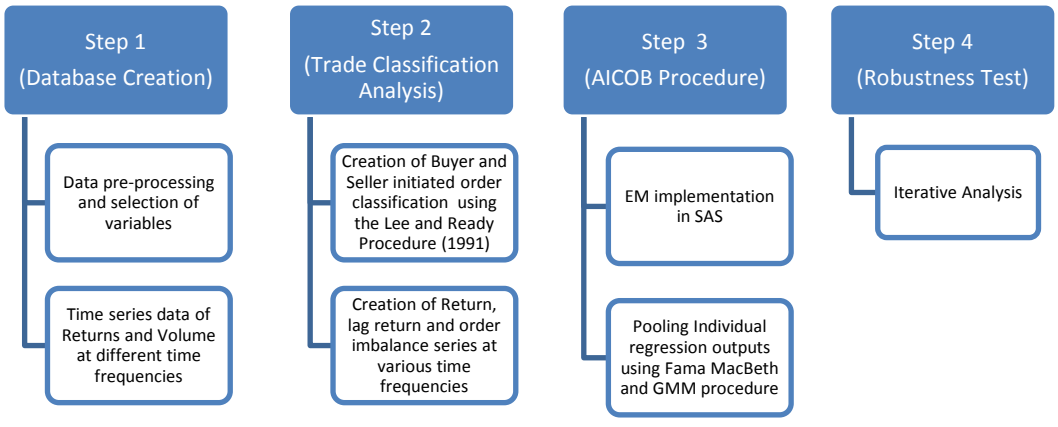


Figure 5.2 Process Diagram of Various Steps in the Data Analysis

5.6 Data Distribution Statistics

Table 5.3 reports the distribution statistics of the sample stocks. As reported in the table, out of 202 firms of the Australian securities exchange, there are 50, 102, and 50 firms in Large,[39] Medium[40] and Small[41] capitalization groups, respectively.

Table 5.3 Summary Statistics of Size-based Groups

Size Classification	Number of Firms	Average Trade Value ($AUD)	Average Trade Size	No. of Orders
Large Cap	50	10,413.05	622.02	1,574,038
Small Cap	102[42]	1,852.79	860.24	479,407
Medium Cap	50	3,573.91	2,057.95	772,275
Average		5279.92	1180.07	
Total	202[43]			2,825,720

[39] Large firms: S&P/ASX 50 represents the large-cap universe for the Australian equity market and is comprised of 50 largest stocks by float adjusted market capitalization.

[40] Medium firms: S&P/ASX 100 is comprised of the 100 largest index-eligible stocks (both large and medium stocks) listed on the ASX by float adjusted market cap. Firms in S&P/ASX 100 but not in S&P/ASX 50 constitute medium firms.

[41] Small Firms: S&P/ASX 200 measures the performance of 200 largest index-eligible stocks (both large and medium stocks) listed on the ASX by float adjusted market cap. Firms in S&P/ASX 200 but not in S&P/ASX 50, S&P/ASX 100, constitute small firms.

www.asx.com.au contains more information on market capitalization.

[42] S&P 200 and S&P 100 have 98 firms in common. Therefore, small cap has 102 firms.

[43] As Small Cap shows 2 extra firms, the total here is not 200 but it is 202.

The average trade value of large-cap stocks is almost five times higher than small stocks. This is consistent with the idea that large-cap stocks are frequently traded with higher potential memory dependent[44] OIB. In terms of trade size, the average trade size (number of trades) of medium stocks is largest among the three groups at 2,057.95 stocks per trade. The table also shows that the average trade value of the market is around $5,000 AUD with a trade size of around 1,000 stocks per trade. This indicates that the average stock price of our sample Australian market is around $5 AUD. As per the table there are around 3 million data points in the dataset. Consistent with Chordia and Subrahmanyam (2004), large-cap stocks have almost three times more data points than small-cap stocks.

5.7 Evaluation

This section contains the experiments conducted using the Australian stock market data to address our research questions. In the first part, we test Chordia and Subrahmanyam (2004) theoretical proposition on the positive relationship between OIB and future returns for individual stocks, by using the largest Australian listed stock (BHP). Later, we extended our experiments to portfolio analysis (multiple stocks). This section describes the mechanics of the experiments conducted but the actual results are presented in the Section 5.8.

[44] Stocks that have high liquidity, receive (from traders) several competing orders that are very close to each other. They share same properties and behavior. Hence, they are interdependent.

5.7.1 Simulation with BHP

We start the empirical analysis with a single stock to test the theoretical propositions of Chordia and Subrahmanyam (2004). In addition to this, it helps to test the theoretical prediction of large stocks' predictability. This will allow us to understand the behavior of individual stock's returns for changes in order imbalance and to understand the mechanics of AICOB procedure. We choose BHP Ltd., one of the most actively traded large firms of the Australian stock market. Tables 5.4 and 5.5 report the results based on 5 minutes and 10 minutes time series frequencies respectively. The values in the tables are based on 12 months trade book data of BHP Ltd. The results in both tables are reported in two panels. Panel A reports results for the trade book data whereas Panel B reports results based on AICOB. The results discussed in detail in Section 5.8 indicate that both Lagged return and Lagged OIB are good predictors of future returns for both AICOB and Trade book data.

5.7.2 Portfolio Analysis

Portfolio analysis helps in generalization of the predictive ability of OIB. Also, institutional investors, who are the major traders in the stock market, maintain portfolios. Their main objective is to enhance portfolio level predictability. All the firms are pooled into size based portfolios. ASX provides the constituent stocks in its website. The S&P/ASX 50 index represents the large-cap universe for the Australian equity market and is comprised of 50 largest stocks by float adjusted market capitalization. The S&P/ASX 100 index is comprised of the 100 largest index-eligible stocks (both large and medium stocks) listed on the ASX by float adjusted market cap.

Firms in the S&P/ASX 100 index but not in the S&P/ASX 50 index constitute medium firms. The S&P/ASX 200 index consists of the 200 largest index-eligible stocks (both large and medium stocks) listed on the ASX by float adjusted market cap. Firms in the S&P/ASX 200 index but not in the S&P/ASX 100 index constitute small firms. Total firms in the sample are divided into three major groups as large-cap, medium-cap and small-cap firms based on their market capitalization as of December 2011. Later, we filter for illiquid stocks or stocks that are not frequently traded and also stocks that have data errors.

Robustness of the Estimation Procedure: Pooled Fama MacBeth GMM estimates

Dataset that contains pooled time series and cross sectional data have higher likelihood of resulting spurious cross-sectional and time series correlations between and within stocks. This results in over estimation bias. The Fama MacBeth (1973) procedure is a popular method in finance literature to address such issues. In this procedure, in the first stage, all regressions are run at individual stock or firm level to obtain the lagged OIB estimated coefficients (individual beta coefficients); thus, avoiding possible cross-sectional correlations between stocks (unlike standard pooled regression procedure). In the second stage, stocks are formed into time base portfolios (monthly portfolios) and regressions are run for portfolios.

However, the Fama MacBeth procedure corrects only cross sectional correlations across all stocks for a particular month. For correcting standard errors of time series regressions, we need to run Generalized Methods of Moments (GMM) method for the time series analysis (Hansen, 1982). We use GMM to regress the time-series estimates

on a constant,[45] which is equivalent to taking a mean. This will adjust standard errors for any possible bias due to time series correlations or auto correlations. All results are based on *Fama MacBeth, GMM* procedure.

5.7.3 Accuracy Dimension

The accuracy based experiments measure how close the AICOB based estimates are compared to the complete order book based estimates. In this dimension, we compare results based on the AICOB estimates with both the trade book estimates and complete order book estimates. All the results are compared across the same time window. We use 5 minutes, 10 minutes and 30 minutes time windows (Chordia et al., 2005). We report results based on portfolio of all stocks for these dimensions. For reporting purposes, the comparison is made by using t-values of the regression coefficients for lagged order imbalance, the proposed major predictor of returns.[46]

5.7.4 Efficiency Dimension

Efficiency dimension measures how long the Lagged OIB can predict returns (in terms of orders placed in 5 minutes, 10 minutes and 30 minutes ahead) by using the AICOB algorithm as against the ordinary trade book based estimates. The longer the predictive ability, the higher is the likelihood of profit trading. Hence, higher efficiency in this context implies more economic efficiency of the algorithm. Given that all the data are

[45] This implies that the portfolio will have zero-beta excess return (no excess return) as expected return is assumed to be the mean return. For more information, refer to Cochrane (2001).

[46] It should be noted that, following existing papers in finance literature (Chordia et al., 2002, 2005 and 2009) all regression models contain lagged returns as the control variable. Also, we do not report the coefficients and other statistics as it takes lot of space and hard to report in a comparative analysis table.

constructed on time interval dimension, efficiency remains the underlying dimension for all results. Consistent with the existing finance literature, time interval is classified into 5 minutes, 10 minutes and 30 minutes intervals, respectively (Chordia et al., 2005).

5.7.5 Adaptability Dimension

Adaptability is measured in three categories. We study the behavior for different firms that are grouped under different sizes (in terms of their market capitalization). Following the rationale used in Section 5.3.1 that, predictions based on lagged OIB varies for different size firms, we measure adaptability under firm size by classifying all stocks into Large Capitalization, Medium Capitalization and Small Capitalization stocks. We expect that the AICOB based order imbalance estimates should be able to predict the returns across all size groups better than trade book based estimates. However, finance theory predicts large capitalization firms to have more significance due to higher frequency of trading activity.

5.8 Results

The results are shown in this section in terms of t-values and its statistical significance. The t-values are reported in parentheses and are shown with *, **, ***, that represent significance at 90%, 95%, and 99% confidence level respectively. The value above the parenthesis is the coefficient of the respective determinants of return (intercept, return lagged OIB). To reject the null hypothesis, the reported t-values in the tables should be more than or equal to the critical threshold[47] values.

[47] t-values greater than equal to 1.645, 1.96 and 2.58 represent significance at 90%, 95%, and 99% confidence level respectively and are the critical threshold values.

5.8.1 Simulation with BHP

Chordia and Subrahmanyam's theoretical propositions are based on Single large stock. However, they report average results of the individual stocks, not single stock per se. Therefore, we start with an individual stock, BHP for better tractability. The results are reported in Table 5.4. The independent variables are lagged OIB and lagged returns. Change in prices (returns) is endogenous to order flow. The table reports that BHP trades for 12 months are categorized into 15,455, 5 minutes intervals. All trades are aggregated for every 5 minutes. Likewise, as per Table 5.5 there are 7838 10 minutes aggregated trades.

Table 5.4 BHP (2012) Results for 5 Minutes

Panel A: Trade book Data Results	
Intercept	-0.006 (-0.69)
Lagged Return	-0.078 (-10.41) ***
Lagged OIB	0.001 (8.96) ***

Panel B: AICOB Data Results	
Intercept	-0.001 (-0.682)
Lagged Return	-0.077 (-10.33) ***
Lagged OIB	0.001 (8.54) ***

This table reports trade book data results and AICOB data results by using lagged OIB, lagged return as independent variables. Lagged OIB is defined as the difference between aggregated buy and sell orders for a defined 5 minutes time interval; t-values are reported in parenthesis below the coefficient; t-values are shown with *, **, *** and represent significance at 90%, 95%, and 99% confidence level respectively.

The results in Table 5.4 indicate that both Lagged return and Lagged OIB are good predictors of future returns. The sign of OIB coefficient is positive and significant (as shown with asterisks; the values that are less than critical threshold value, are not significant and hence, do not have asterisks). This indicates that future returns move in the same direction as their past order imbalance. The finance literature (Chordia et al.,

2005) predicts positive and significant coefficient for Lagged OIB variable. This implies that a positive order imbalance (where buys are more than sells) leads to increase in stock price.

Table 5.5 BHP (2012) Results for 10 Minutes

Panel A: Trade book Data Results	
Intercept	-0.005 (-0.82)
Lagged Return	-0.053 (-4.82) ***
Lagged OIB	0.001 (4.21) ***

Panel B: AICOB Data Results	
Intercept	-0.005 (-0.69)
Lagged Return	-0.051 (-4.58) ***
Lagged OIB	0.001 (3.4) ***

This table reports trade book data results and AICOB data results by using lagged OIB, lagged return as independent variables. Lagged OIB is defined as the difference between aggregated buy and sell orders for a defined 10 minutes time interval; t-values are reported in parenthesis below the coefficient; t-values are shown with *, **, *** and represent significance at 90%, 95%, and 99% confidence level respectively.

Order imbalance is defined as the difference between buy and sell orders, hence, positive order imbalance implies more buy orders compared to sell orders. This implies, return increases with higher buy orders. Panel B reports results based on AICOB procedure. We run AICOB procedure on the full dataset of trade book data, however, by randomly missing lagged order imbalance values. These values are imputed through EM method by observing lagged returns (as discussed in Section 4.6). The AICOB based regression results indicate that the results are very close (in terms of coefficients and t-values) to the full dataset results of the trade book.

The results in Table 5.5 for 10 minutes time intervals are similar to Table 5.4 results. In summary, the simulation, that compares trade book data (Chordia Subrahmanyam, 2004) with AICOB data, shows that AICOB based estimates give

consistent results with the literature. We now focus on portfolio analysis for more robust and general results on prediction.

5.8.2 Portfolio Analysis under Accuracy and Efficiency Dimensions

Table 5.6 reports pooled regression estimates based on the Fama MacBeth, GMM method for all the Australian listed stocks. The objective of this analysis is to assess whether the AICOB-based results are superior to trade book based results and also whether the AICOB-based results are consistent with the complete order book results.

We combine both accuracy and efficiency dimension based results due to the fact that all data are arranged in regular time intervals. Hence, for all results, efficiency remains the underlying evaluation dimension. We report adaptability results separately as there are sub divisions in the adaptability analysis. The first set of results compare results based on the AICOB under accuracy and efficiency dimensions. The results in Table 5.6 are arranged in three panels from Panel A to Panel C. Panel A reports results based on trade book observations.

Panel B reports results based on the AIOCB procedure, where several trade book observations are replaced with imputed estimates. Panel C reports results based on complete order book observations that entered into the order book system. Accuracy is determined based on comparison between AICOB results in Panel B and complete order book based results in Panel C. As described in Section 5.2.2, we measure efficiency in terms of how long order imbalance prediction lasts.

Table 5.6 Accuracy and Efficiency Analysis for All Stocks

Variable Name	5 Minutes	10 Minutes	30 Minutes
Panel A: Trade Book data			
Intercept	-0.19	0.16	-0.12
(t-value)	(-2.12)**	(2.15)**	(-1.8)**
Lagged OIB	0.01	0.01	0.01
(t-value)	(1.79)*	(2.18)**	(1.04)
Lagged Return	-0.30	-0.33	-0.38
(t-value)	(-27.82)***	(-19.8)***	(-22.65)***
Adjusted R-Square	0.11	0.14	0.22
No. of Observations	119625	86103	29615
Panel B: AICOB data			
Intercept	1.94	0.16	0.12
(t-value)	(2.10)**	(2.18)**	(1.8)**
Lagged OIB	0.01	0.01	0.01
(t-value)	(1.75)*	(1.14)	(0.83)
Lagged Return	-0.30	-0.33	-0.38
(t-value)	(-27.87)***	(-19.88)***	(-22.79)***
Adjusted R-Square	0.11	0.14	0.22
No. of Observations	119625	86103	29615
Panel C: Order book data			
Intercept	-0.80	-0.98	-0.81
(t-value)	(-2.01)**	(-10.92)***	(-6.20)***
Lagged OIB	0.01	0.01	0.01
(t-value)	(1.86)**	(0.53)	(1.17)
Lagged Return	-0.33	-0.19	-0.15
(t-value)	(-2.41)***	(-11.10)***	(-2.53)***
Adjusted R-Square	0.07	0.07	0.07
No. of Observations	234724	184986	85579

This table reports trade book data result, AICOB data results and order book data results by using lagged OIB, lagged return as independent variables for All Stocks. Lagged OIB is defined as the difference between aggregated buy and sell orders for 5 minutes, 10 minutes time and 30 minutes intervals; t-values are reported in parenthesis below the coefficient; t-values are shown with *, **, *** and represent significance at 90%, 95%, and 99% confidence level respectively.

All orders and trades are aggregated at 5, 10 and 30 minutes intervals. The number of intervals varies in each panel based on the recorded observations during those intervals. Panel C in Figure 5.6 that has order book data, has significantly more observations (471,904 compared to 119,625) due to higher orders compared to trades in the stock market. Lagged OIB represented one interval lag value of order imbalance

between buy and sell orders aggregated at 5, 10 and 30 minutes intervals. Panel A results (t-value of lagged OIB) suggest that, lagged order imbalance can predict stock returns for 5 and 10 minutes intervals. This is evident from t-values of Lagged OIB coefficients being 1.79 and 2.18 indicating significance at 90% and 95% confidence level respectively. However, Panel B results do not support the hypothesis that Lagged OIB values based on AICOB procedure are superior to Panel A based results. The t-value of the Lagged OIB estimates in Panel B is not significant beyond 5 minutes interval, as the values for 10 minutes and 30 minutes duration windows are below the critical threshold and thus does not reject the null hypothesis.

From the Table 5.6 it can be noticed that the adjusted R-squared[48] of the complete order book estimates is lower than that of trade book and AICOB estimates. In other words, many data points are away from the regression line. This suggests that the complete order book data of the pooled sample would have lot of variation in orders placed. Given that the stocks range for very small to very large; several data points would be away from the regression line. In the case of trade book data, the data points would be closer to the mean as trade book contains only matched orders. Several unmatched orders of complete order book would lead to lower adjusted R-squared. In addition to that, in pooled datasets, that have both cross sectional and time series variance, adjusted R-Squared is difficult to interpret as it contains both cross sectional and time series variation. Hence, running regressions separately for different firm size

[48] Coefficient of determination that explains how close the data are to the fitted regression line.

based groups would improve the adjusted R-Squared. Our next analysis focuses on adaptability dimension where we report results on firm size based groups.

5.8.3 Portfolio Analysis under Adaptability Dimension

We test the adaptability of the AICOB for various firm sizes. As predicted by the literature, larger firms with more trading activity tend to have higher predictive ability compared to smaller firms. As reported in Table 5.7, we divided all firms into three groups based on their market capitalization terciles as small, medium and large firms. Table 5.7 results are pooled regression estimates based on Fama Macbeth GMM procedure for all observations segregated into large-cap or capitalization, medium-cap and small-cap firms. Firm size is measured using market capitalization. The purpose of this table is to understand whether the predictive power of order imbalance varies based on firm size. Accuracy of the results is determined based on comparison between results in Panel B and Panel C. All observations are aggregated at 5, 10 and 30 minutes intervals, respectively. Lagged OIB represented one interval lag value of order imbalance between buy and sell orders aggregated at 5, 10 and 30 minutes intervals. In Table 5.7, Panel A, which reports results based on Trade book data, shows Lagged OIB is significant for mid-cap stocks at 5 minutes interval.

The predictive power of Lagged OIB estimates for large firms and small firms for all other intervals is not significant. Panel B results, based on AICOB procedure, find significant predictive power for large-cap firms, for 5 minutes intervals. R-Squared value for the large-cap stocks in Panel C in the 5 minutes interval stands at 0.49 and drops for the 10 minutes and 30 minutes intervals for the same capitalization stocks. One

of the main reasons for such drop is that the number of observations decreases with the increase in the time period. This is much more significant for large stock compared to others. This is one of the plausible explanations for drop in the R-Squared value. This result is more consistent with the extant literature on order imbalance prediction of future returns. More importantly, Panel B, AICOB results are consistent with Panel C, complete order book results. This shows that AICOB based lagged OIB can significantly determine contemporaneous returns. It should be noted that the adjusted R-Squared of Panel C are lower in few cases compared to Panel A and Panel B. In fact Panels A and B adjusted R-Squared are much closer. This is due to similar orders (mainly matched orders) in both Panel A and Panel B. In the case of complete order book, as it contains both matched and unmatched orders, the overall unexplained variance (that drives the adjusted R-Squared numbers) increases.

In the case of Panels A and B that contained matched orders (which have relatively higher influence on returns) the unexplained variance will be relatively lower. Compared to the adjusted R-Squared estimates of Panel C in Table 5.6, the Panel C estimates of Table 5.7 are much higher. This again confirms the complexity associated with pooling of firms of several sizes that tend to have significant variation in the orders placed by traders. Consistent with the literature, we find that lagged returns are negatively correlated with contemporaneous returns (Chordia and Subrahmanyam, 2004). In summary, the results in Table 5.6 and Table 5.7 confirm that AICOB predictions are more accurate (as they are closer to order book) and consistent with order book results when compared to trade book based estimates that suffer from missing data problem.

Table 5.7 Adaptability Analysis

Variable Name	Large Cap (5 minutes)	Mid Cap (5 minutes)	Small Cap (5 minutes)	Large Cap (10 minutes)	Mid Cap (10 minutes)	Small Cap (10 minutes)	Large Cap (30 minutes)	Mid Cap (30 minutes)	Small Cap (30 minutes)
Panel A: Trade Book data									
Intercept	-0.14	0.07	0.01	-0.23	0.04	0.03	0.03	-0.02	-0.03
(t-value)	(-0.75)	(0.49)	(0.14)	(-1.29)	(0.29)	(0.45)	(0.95)	(-0.86)	(-1.15)
Lagged OIB	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
(t-value)	(1.18)	(1.74)*	(1.28)	(1.56)	(1.01)	(1.52)	(0.83)	(0.96)	(1.31)
Lagged Return	-0.42	-0.42	-0.46	-0.49	-0.51	-0.59	-0.92	-0.96	-0.78
(t-value)	(-5.78)***	(-4.83)***	(-13.99)***	(-7.46)***	(-5.11)***	(-11.28)***	(-14.24)***	(-46.72)***	(-13.50)***
Adjusted R-Square	0.25	0.25	0.26	0.32	0.35	0.41	0.91	0.94	0.69
No. of Observations	39650	51850	28125	23288	26095	36720	7560	9095	12960
Panel B: AICOB data									
Intercept	0.14	0.07	0.01	0.23	0.04	0.03	0.03	0.02	0.03
(t-value)	(0.75)	(0.49)	(0.16)	(1.28)	(0.28)	(0.44)	(0.95)	(0.93)	(1.16)
Lagged OIB	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
(t-value)	(2.10)**	(0.07)	(0.11)	(2.81)***	(2.06)**	(1.75)*	(0.1)	(0.85)	(0.68)
Lagged Return	-0.43	-0.42	-0.46	-0.49	-0.51	-0.59	-0.92	-0.96	-0.78
(t-value)	(-5.81)***	(-4.83)***	(-13.92)***	(-7.46)***	(-5.17)***	(-11.34)***	(-14.24)***	(-45.78)***	(-13.59)***
Adjusted R-Square	0.25	0.25	0.25	0.31	0.35	0.41	0.91	0.94	0.69
No. of Observations	39650	51850	28125	23288	26095	36720	7560	9095	12960
Panel C: Order book data									
Intercept	-0.07	-0.04	0.01	-0.07	-0.51	0.01	-0.01	0.24	-0.03
(t-value)	(-0.52)	(-0.27)	(0.24)	(-0.52)	(-0.29)	(0.01)	(0.06)	(1.11)	(-0.84)
Lagged OIB	0.01	0.01	0.01	0.01	0.01	0.01	0.01	.01	0.01
(t-value)	(2.88)***	(0.4)	(1.43)	(2.79)***	(2.3)**	(2.67)***	(4.2)***	1.08)	(1.66)
Lagged Return	-0.23	-0.03	-0.57	-0.23	-0.36	-0.62	-0.48	-0.42	-0.69
(t-value)	(-7.43)***	(-3.70)***	(-10.31)***	(-7.42)***	(-3.28)***	(-12.09)***	(-5.55)***	(-6.46)***	(-10.39)***
Adjusted R-Square	0.49	0.13	0.34	0.04	0.14	0.34	0.09	0.14	0.40
No. of Observations	46534	56474	78716	46534	29536	78716	9073	10977	15359

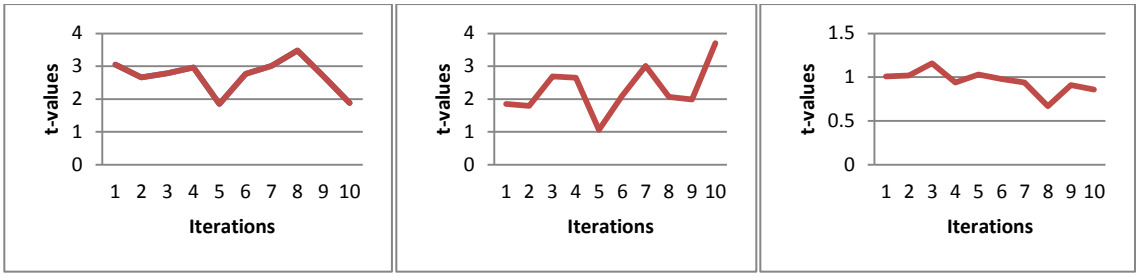
This table reports trade book, AICOB and order book data results by using lagged OIB, lagged return as independent variables for Large, Medium and Small firms. Lagged OIB is defined as the difference between aggregated buy and sell orders for 5 minutes, 10 minutes time and 30 minutes intervals; t-values are reported in parenthesis below the coefficient; t-values are shown with *, **, *** and represent significance at 90%, 95%, and 99% confidence level, respectively.

5.8.4 Robustness Check

The results in the previous section indicate that AICOB is beneficial in predicting returns with Lagged OIB estimates for large-cap stocks and mainly up to 5 minutes lag time interval. To further show that results for large-cap stocks are not significant just by chance, we test if the results are subject to potential sample selection bias. We check whether the AICOB based results are robust to any random sample of portfolios where, for each iteration, stocks are selected randomly.

If the results are consistent and significant across all random samples, then the results in the above sections are treated as robust for sample selection bias. We ran 10 iterations of AICOB based results for each of the three time intervals (5, 10, 30 minutes intervals), each time, changing the sample of large, medium and small-cap stocks randomly to see whether the results reported in our adaptability analysis are consistent across different iterations.

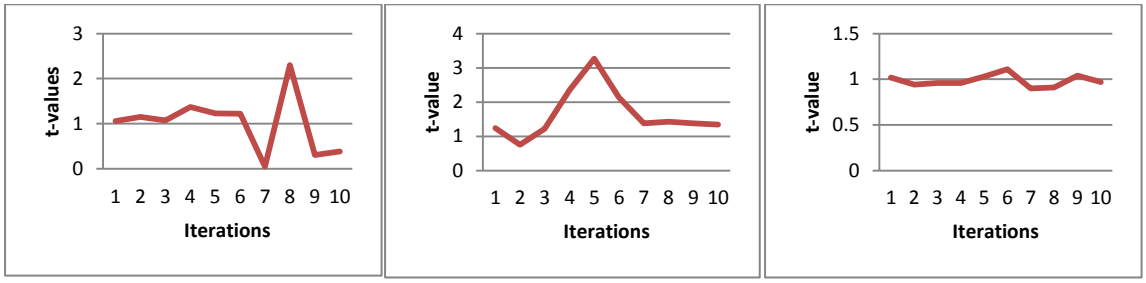
The purpose of this analysis is to understand whether OIB estimates are stable for randomly chosen data of large-cap, medium-cap and small-cap firms at 5, 10 and 30 minutes respectively. We report the t-values of lagged order imbalance coefficients using our AICOB procedure for 10 random sample iterations. Figure 5.3, Figure 5.4 and Figure 5.5 report graphs of iterations and t-values of order imbalance coefficients on X-axis and Y-axis respectively.



a) LC 5 Mins **b) LC 10 Mins** **c) LC 30 Mins**

Figure 5.3 Robustness of the AICOB for Large-Cap Stocks

The AICOB based iterative results indicate that for 5 intervals of large-cap stocks, irrespective of the random sample, the results consistently indicate that lagged order imbalance can predict contemporaneous returns as the t-values for almost all the iterations is above the threshold value of 1.96.As can be seen, for the results for 30 minutes interval, none of the iterations are significant (t-values less than 1.645).



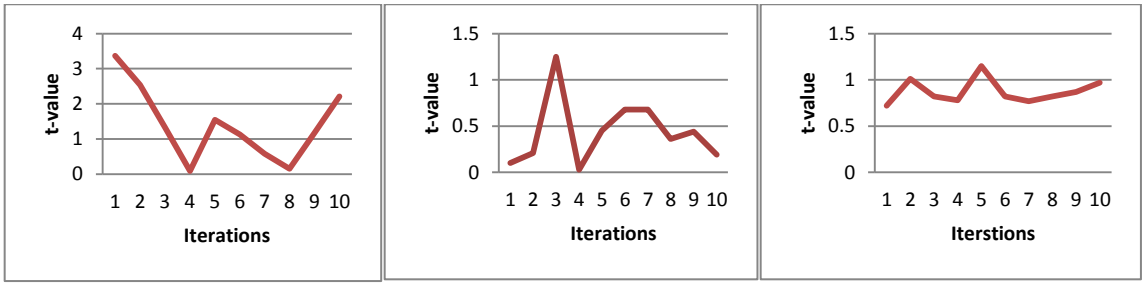
a) MC 5 Mins **b) MC 10 Mins** **c) MC 30 Mins**

Figures 5.4 Robustness of the AICOB for Medium-Cap Stocks

In the case of 10 minutes intervals, the results show variation where some iterations are significant and some are not. The results in Figure 5.4 and Figure 5.5 report iterations based t-values for medium and small-cap stocks for 5, 10 and 30 minutes intervals, respectively. These results indicate that predictive power of lagged

order imbalance is quite erratic in various iterations. For instance, for small-cap 5 minutes intervals, in the Figure 5.5-a, some are highly significant and some are not.

The results in Figure 5.5-a, indicate that in some iterations, there is a possibility that lagged order imbalance can predict returns (as the t-value for some is greater than equal to 1.645). There is no consistency with change within the iterations. In few cases, the results seem to be significant (as shown in Figure 5.4-b where t-value is greater than 3.0 for iteration 5). It is purely by chance that in some cases OIB can predict for mid-cap stocks. Likewise, there is a possibility that the significant prediction of trade book based estimate for medium stocks for 5 minutes in Table 5.7 could have been just by chance. Hence, there is every chance that trade book data based results for medium and small-cap firms are not reliable.



a) SC 5 Mins **b) SC 10 Mins** **c) SC 30 Mins**

Figures 5.5 Robustness of the AICOB for Small-Cap Stocks

In summary, prediction based on order imbalance could be quite sensitive to missing data problem. The EM algorithm imputation process is a machine learning procedure that is aimed at imputing the missing values. The imputation of missing values is more reliable for large stocks that have frequent trading and hence few missing observations can be relatively more accurately replaced with imputed values. The

rationale behind this is that as the number of missing values increases, the estimation power of EM algorithm will decrease. Hence, EM estimation works better for relatively lower missing values. Given that large stocks have lower missing values (due to higher trading activity), the prediction power of EM algorithm will be better for large stocks compared to medium and small stocks. The large number of memory dependent orders ensures better estimation and corresponding imputation of missing orders information. Hence, the justification is not based on the economic rationale. It is mainly based on the data sensitivity issue in the estimation procedure.

5.9 Interpretation of the Results

For large stocks with more history dependent orders, missing data problem do not affect prediction of future returns. As is evident from Table 5.4 and Table 5.5, for BHP stock, trade book's as well as AICOB's results are significant. However, as the firm size decreases and when firms are treated as a portfolio, the complexities associated with missing data problem become severe. Significant prediction by trade book for a single large stock therefore, cannot be used to draw any conclusions without studying the complexities associated with missing data problem in portfolios. Reporting BHP Ltd. results will provide a logical path for portfolio results.

For portfolio of all stocks in Table 5.6, it is evident from the t-values that Lagged OIB coefficients significantly predict returns for all stocks, for 5 minutes and 10 minutes intervals using trade book data. AICOB and order book does not show any significance in their t-values for the 10 minutes intervals. As discussed in goodness of the prediction section (Section 5.3), complete order book data are the true benchmark for comparison.

When we compare AICOB's and order book's results, we find a lot of similarities as both report no significant relationship between lagged order imbalance and contemporaneous returns for beyond 5 minutes interval. Whereas, trade book based predictions are quite different from the complete order book data. Evaluation against the actual order book data is a better procedure and the results are superior in this context.

In Table 5.7, which reports firm size, portfolio results, trade book data shows Lagged OIB is not significant except for mid-cap stocks at 5 minutes interval. This suggests that the trade book result of all stocks portfolio, of Table 5.6, for 5 minutes and 10 minutes intervals are driven by the medium stocks. This result is not consistent with the theoretical predictions of Chordia and Subrhamanyam (2004) that indicate that the predictive power should be higher for large-cap stocks with more trading activity and higher memory-dependent orders. Had the trade book been a good predictor, then the results should have been significant for large-cap. As they are not significant, it goes to show that the trade book is not a good predictor of future returns.

The results based on both AICOB and complete order book indicate that lagged order imbalance is a significant predictor of contemporaneous returns for large-cap 5 minutes and 10 minutes intervals. Also, the results are significant for 10 minutes intervals of Medium Cap and Small Cap firms. In the case of trade book, however, the results are significant only for Medium cap firms, for only 5 minutes interval. These results indicate that AICOB results are closer to the complete order book results compared to the trade book results. Hence, these results further support the conjecture that trade book (with missing data) is insufficient to predict future returns.

5.10 Summary

In this chapter, we test our AICOB procedure with a pre-determined and structured evaluation strategy. Our main objective is to find whether addressing missing data problem with machine learning methods improve stock return prediction. Hence, we evaluate the developed methodology by measuring: (1) how accurate is the prediction based on the proposed methodology (relative to the non-missing complete data based predictions); (2) how efficient is the methodology (in terms of the duration of prediction); and (3) how adaptable is the methodology (in terms of varying firm size).

A four-step implementation strategy of AICOB procedure is presented, where the required parameters that are used to create OIB estimates using both trade book data and complete order book data are computed. For AICOB based OIB estimates, the methodology proposed in Chapter 4 is used. A robust regression testing methodology by using Fama MacBeth and GMM procedures is also proposed in this chapter. Later, we use the Australian stock market data for executing and analysing results within the guidelines of the proposed evaluation strategy. It is found that AICOB based OIB estimates predict returns better than trade book based estimates. The AICOB passes accuracy test and adaptability test with additional robustness analysis. In terms of efficiency, it is found that AICOB is efficient for 5 minutes interval duration. For trading purposes, 5 minutes is a long trading window for engaging in profitable trading activity. The results indicate AICOB, which is expected to address the missing data problem, serves its purpose. One new implication of our results is that the existing notion of generalizability of OIB being a robust predictor of future returns, for all stocks needs

further examination in the light of these empirical results supported by complete order book data.

In summary, this chapter develops and tests an objective evaluation strategy for AICOB, based on efficiency, accuracy and adaptability dimensions. The thesis uses Australian stock market data that provides not only trade book data but also historical order book data for cross validating the results. The main achievement is that AICOB based predictions match with the complete order book data. Whereas, trade book based predictions are quite different and inconsistent compared to the complete order book data.

The AICOB based results are also consistent with the theoretical predictions proposed in finance literature that OIB predicts better as the firm size increases. The results show that large firms, with higher trading activity and more competition for order flows, reports more significant OIB prediction of future returns. Trade Book based OIB estimates, which suffer from missing data problem, fail to predict future returns for stock portfolios. Hence, addressing the missing data problem is important before implementing OIB based trading strategies.

This page is intentionally left blank.

Chapter 6

CONCLUSION

6.1 Research Summary

Missing data problem is ubiquitous in many real life situations. Information Technology researchers have explored and tried to address this problem in different settings. In this thesis, we undertake research to address missing data problem associated with order book information in stock markets. This is an in depth and large-scale study with systematic and comprehensive framework to address missing data problem in the finance literature.

Traders place orders that influence price revisions and such possible change in prices is termed as return on investment. Given that predicting future return is the central objective of financial investments, understanding the drivers of price change is the key for financial forecasting. Orders placed by the traders are the primary drivers for such temporal price changes and corresponding returns on investments (Chordia and Subrahmanyam, 2004). However, as highlighted in the first chapter of the thesis, transparency regulation imposed in many stock exchanges restricts traders to exploit sensitive order book data on a real time basis. Majority of the stock exchanges reveal only matched orders reported through trade book. Even this information is reported with a delay.

Trade book, that is reported and visible to traders, attracts missing data problem as many orders placed by traders, that are unmatched, are not transferred from the order book. Such unmatched orders would still influence future returns, however, they are not accounted for, while calculating order imbalance using trade book data. Hence, order imbalance, calculated from trade book, being incomplete, obstructs investors' ability to forecast future returns accurately. The thesis tries to address this problem with a multi-disciplinary approach using machine learning for stock market data.

The thesis contributes by attempting to integrate advances in Information Technology research with the advances in Finance research and thus paves new path for a multi-disciplinary research framework. This is achieved by developing an integrated theoretical framework relating to a finance missing data problem with a machine based adaptive learning methodology. The thesis is motivated by illustrating the importance of order imbalance in predicting future returns.

The missing data problem in order imbalance calculation is addressed by theoretical representation of the missing data problem as a Missing at Random data and estimating log likelihood functions for single and joint variables within the context of missing data problem. This foundation for understanding the mechanics of the proposed estimation procedure provides the first insight that the log likelihood values for incomplete data (OIB from trade book) favors a lower value for OIB mean due to incomplete data records. Maximum likelihood estimation improves the accuracy of the parameter estimates by borrowing information from the incomplete records of the observed orders.

Hence, such downward adjustment to the OIB average effectively steers the estimator towards an OIB mean for missing data that is identical to that of complete data. A running example (as shown throughout Chapter 3) using an Australian stock, AGK (AGL Energy Ltd.) explains the theoretical intuition. The example demonstrates that estimating OIB by using incomplete order records, as against deleting missing information of unmatched orders, improves the accuracy of the parameter estimates.

The thesis proposes a Relational Markov Network (RMN) model for estimation of the joint distribution function of orders, order characteristics and their interactions. An adaptation of the Expectation Maximization Algorithm is proposed to address the missing data problem during the joint estimation procedure. The proposed novel methodology overcomes the estimation problem, especially in the context of missing order book data. This is in contrast to that of regression models that do not identify the relationship between order imbalance and returns. Orders and their imbalance at the given point of time influence the price movements (and the corresponding returns) and such interdependencies, if not jointly estimated, lead to biased estimates. RMN framework allows us to overcome this problem by establishing the relational network between variables and further establishing joint probabilistic functions. EM algorithm then maximizes the likelihoods of the estimates by learning information related to the missing orders through posterior distribution functions and thus improving the accuracy of the estimates. The proposed algorithm is termed as Algorithm for Imputed Complete Order Book (AICOB).

AICOB has been evaluated on efficiency, accuracy and adaptability dimensions. The thesis uses the Australian stock market data that provide not only trade book data

but also historical order book data for its evaluation. This provides a unique setting to cross validate the accuracy of AICOB methodology. The results based on the significance of t-values of the OIB coefficients, show that AICOB based predictions are much closer to the complete order book data. Whereas, trade book based predictions are quite different and inconsistent from the complete order book data. The AICOB based results also support the earlier claim by Chordia et al., 2004 that OIB becomes a good predicator of future returns as the firm size increases. The results show that large firms, with higher trading activity and more competition for order flows, report more significant OIB prediction of future returns. Hence, addressing the missing data problem is important before implementing OIB based trading strategies.

6.2 Research Contributions

The major contributions of this thesis can be summarized as follows:

- *To develop a systematic framework for addressing the missing data problem in stocks returns' forecasting*

  The thesis models the missing data problem as a Missing at Random (MAR) data and builds a systematic framework to estimate single as well as joint log likelihood functions. In the thesis it is shown that estimating using incomplete records, improves the accuracy of the parameter estimations. The results support the argument that machine learning based tools aid in predicting of future returns, especially, when the analysis involves dealing with missing data problem. From this perspective, AICOB (Algorithm for Imputed Complete Order Book), which offers a comprehensive procedure, is implemented. AICOB shows consistent results to

predict future returns for stock portfolios of large firms and some result for medium and small firms. Thus, the thesis contributes through a new and systematic framework to address missing data problem associated with order book.

- *To measure the prediction accuracy through the empirical evaluation strategy and provide a comparative analysis of AICOB procedure*

  The thesis proposes a Relational Markov Network Model of the joint distribution function of orders, order characteristics and their interactions to estimate the expected value or the mean of the distribution. The expected values based on these distributions are used as the initial estimates in the proposed machine learning methodology. By using the initial values, the values for the missing data are re-estimated iteratively through the Expectation Maximization Algorithm (EMA). EMA helps to address the missing data problem during the joint estimation procedure of the portfolio of stocks. All pooled regression results follow *Fama MacBeth* and Generalized Methods of Moments (GMM) procedures. These procedures control the cross sectional and time series correlations between the observations and across the pooled stocks. The proposed novel methodology overcomes the estimation problem and improves prediction accuracy in the context of missing order book data.

  Further, the proposed methodology provides comparative analysis with the actual complete order book data. Both AICOB and trade book based missing data estimates are evaluated against the complete order book data under accuracy, adaptability and efficiency dimensions. A sensitivity analysis through random

sampling procedure confirms that AICOB is a better predictor of future returns based on the significance of the t-values of the OIB coefficients (as shown in the tables, in Chapter 5). In summary, the thesis contributes a new method for addressing the missing data problem while estimating the order imbalance in predicting future returns.

- *Understanding the impact of order book data on the price discovery process*

  From this perspective, the finance literature suggests that, heavily traded stocks with high competition among the traders, should exhibit efficient price discovery process (Fama, 1972), with no dependency on available information. However, the literature starting from research papers like Chordia and Subrahmanyam (2004) suggests that heavily traded stocks, with many traders placing competitive orders that are very close to each other, will have more history dependence. Such history dependence allows orders and their corresponding imbalance to predict temporal or short terms price changes or returns.

  Disentangling this price efficiency debate is one of the contributions of this thesis. The study supports the conjecture and also shows that order imbalance is a significant predictor of future returns for heavily traded large stocks. Unlike, existing research, this claim is supported by complete order book data. An implication of this finding is that the stock exchanges can be lenient with the transparency regulation, mainly for large stock, as sensitive order information is useful for only large and heavily traded stocks that are priced more efficiently.

6.3 Future Research Directions

The research work implemented in this thesis sheds light on different areas of research for further possible extensions. This section highlights some of those key directions as follows:

- *Application of AICOB to other missing data problems in finance*

  Finance datasets in many situations suffer from missing data problem. Missing data are a major hindrance in predicting stock returns accurately. Linnainmaa and Saar (2012) show that broker ID information is important for predicting price impact. The authors find that when broker ID information is not revealed, informed investors would be able to hide their trades more effectively. Hence, anonymity leads to higher profits to the informed investors. Foucault et al. (2007), show that trader anonymity influences trading cost and information content. These papers suggest that missing data in the form of identity of the brokers and traders is highly informative. Future research can focus on this identification based missing data problem as all major stock exchanges, including top three[49] stock exchanges in the world, namely, the New York Stock Exchange, the NASDAQ and the London Stock Exchange do not disclose broker identification.

  Also, assessing credit risk of bank customers involves collecting data on customer profiles. In several instances the datasets will be incomplete. For instance, prospective borrowers may not fill all the details required to assess their credit

[49] Top three stock exchanges according to Forbes (www.forbes.com, last accessed on 31/07/2016).

worthiness. In such situations, AICOB can be implemented to fill the missing data by learning from the incomplete records.

- *Application of AICOB to different stock market architectures*

  Trading rules vary from market to market. Some markets are more transparent than others. The future research can assess the sensitivity of AICOB with varying market transparency level. For instance, Hendershott and Jones (2005) show that when Electronic Communications Network (ECNs are electronic platforms that allow traders to trade on major US stock markets) change the transparency regulation from displaying order book information to banning such information, the price discovery in these markets decline. Implementing AICOB in different market architectures and performing a comparative analysis can help market regulators to frame optimal transparency regulation.

- *Other areas of future research*

  AICOB provides information on the missing orders that are not visible to traders. Hence, with AICOB information we can estimate the hidden orders demand on stock prices. In the future, AICOB can be tested on:

  (i) Efficiency and accuracy pre/post earnings announcement that contain price sensitive information.

  (ii) Check whether hidden orders information varies across domestic stocks versus cross-listed stocks.

  (iii) What is the hidden orders' impact during up/down markets?

  (iv) Test if hidden orders' impact deteriorates/improves as one moves away from an optimum trading price range.

(v) AICOB could be tested for after the announcement of the open market stock buyback programs since the complete order book is more likely to contain company initiated trades which necessarily contain more information.

(vi) Study if ownership matters since stocks primarily held by institutions may suffer less from 'lack of information'.

In conclusion, this thesis takes a step forward and opens up new opportunities in understanding the missing data problem in stock markets by realising the potential of applying Information Technology (IT) research to Finance problems and there by establishing a common ground for cross-disciplinary research in the areas of IT and Finance.

This page is intentionally left blank.

BIBLIOGRAPHY

Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y., 2015. Time-series Clustering–A Decade Review. *Information Systems*, *53*, pp.16-38.

Ahmed, N.K., Atiya, A.F., Gayar, N.E. and El-Shishiny, H., 2010. An Empirical Comparison of Machine Learning Models for Time series Forecasting. *Econometric Reviews*, *29*(5-6), pp.594-621.

Alpaydin, E., 2004. Introduction to Machine Learning. *The MIT Press.*
Barr, D.S. and Mani, G., 1994. Using Neural Nets to Manage Investments. *AI Expert*, *9*(2), pp.16-21.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. *Oxford University Press.*

Breiman, L., Friedman, J., Stone, C.J. and Olshen, R.A., 1984. Classification and Regression Trees. *CRC Press.*

Carling, A., 1992. Introducing Neural Networks. *Sigma Press.*

Chapados, N. and Bengio, Y., 2001. Cost Functions and Model Combination for VaR-based Asset Allocation Using Neural Networks. *IEEE Transactions on Neural Networks*, *12*(4), pp.890-906.

Chiang, W.C., Urban, T.L. and Baldridge, G.W., 1996. A Neural Network Approach to Mutual Fund Net Asset Value Forecasting. *Omega*, *24*(2), pp.205-215.

Chordia, T. and Subrahmanyam, A., 2004. Order Imbalance and Individual Stock Returns: Theory and Evidence. *Journal of Financial Economics*, *72*(3), pp.485-518.

Chordia, T., Roll, R. and Subrahmanyam, A., 2000. Commonality in Liquidity. *Journal of Financial Economics*, *56*(1), pp.3-28.

Chordia, T., Roll, R. and Subrahmanyam, A., 2005. Evidence on the Speed of Convergence to Market Efficiency. *Journal of Financial Economics*, *76*(2), pp.271-292.

Daniel, K., Hirshleifer, D. and Subrahmanyam, A., 1998. Investor Psychology and Security Market Under-and Overreactions. *The Journal of Finance*, *53*(6), pp.1839-1885.

Deb, S.S. and Marisetty, V.B., 2010. Information Content of IPO Grading. *Journal of Banking & Finance*, *34*(9), pp.2294-2305.

Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp.1-38.

Desai, V.S., Crook, J.N. and Overstreet, G.A., 1996. A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. *European Journal of Operational Research*, *95*(1), pp.24-37.

Donaldson, R.G. and Kamstra, M., 1996. Forecast Combining with Neural Networks. *Journal of Forecasting*, *15*(1), pp.49-61.

Fadlalla, A. and Lin, C.H., 2001. An Analysis of the Applications of Neural Networks in Finance. *Interfaces*, *31*(4), pp.112-122.

Fama, E.F. and MacBeth, J.D., 1973. Risk, Return, and Equilibrium: Empirical Tests. *The Journal of Political Economy*, pp.607-636.

Fama, Eugene (1970). Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, 25, pp. 383-417.

Fletcher, D. and Goss, E., 1993. Forecasting with Neural Networks: An Application Using Bankruptcy Data. *Information & Management*, *24*(3), pp.159-167.

Foucault, T., Moinas, S. and Theissen, E., 2007. Does Anonymity Matter in Electronic Limit Order Markets?. *Review of Financial Studies*, *20*(5), pp.1707-1747.

Funahashi, K.I., 1989. On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks*, *2*(3), pp.183-192.

Gales, M. and Young, S., 2008. The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, *1*(3), pp.195-304.

Gelman, A. and Hill, J., 2006. Data Analysis using Regression and Multilevel/Hierarchical Models. *Cambridge University Press*.

Ghahramani, Z. and Jordan, M.I., 1997. Factorial Hidden Markov Models. *Machine Learning*, *29*(2-3), pp.245-273.

Glorfeld, L.W. and Hardgrave, B.C., 1996. An Improved Method for Developing Neural Networks: The Case of Evaluating Commercial Loan Creditworthiness. *Computers & Operations Research*, *23*(10), pp.933-944.

Gonzales, R.C. and Woods, R.E., Digital Image Processing. 2002. *Prentice Hall*.

Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. *Journal of econometrics*, *45*(1), pp.39-70.

Hansen, L.P., 1982. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica: Journal of the Econometric Society*, pp.1029-1054.

Hardle, W., 1990. Applied Nonparametric Regression. *Cambridge University Press*.

Harris, L.E. and Panchapagesan, V., 2005. The information content of the limit order book: evidence from NYSE specialist trading decisions. *Journal of Financial Markets*, *8*(1), pp.25-67.

Haykin, S., 1994. Neural Networks: A Comprehensive Foundation. *Macmillan Publishers.*

Hendershott, T. and Jones, C.M., 2005. Trade-through Prohibitions and Market Quality. *Journal of Financial Markets*, *8*(1), pp.1-23.

Hendershott, T., Jones, C.M. and Menkveld, A.J., 2011. Does Algorithmic Trading Improve Liquidity?. *The Journal of Finance*, *66*(1), pp.1-33.

Hornik, K., Stinchcombe, M. and White, H., 1989. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, *2*(5), pp.359-366.

Hung, S.Y., Liang, T.P. and Liu, V.W.C., 1996. Integrating Arbitrage Pricing Theory and Artificial Neural Networks to Support Portfolio Management. *Decision Support Systems*, *18*(3), pp.301-316.

Imandoust, S.B. and Bolandraftar, M., 2013. Application of k-Nearest Neighbor (kNN) Approach for Predicting Economic Events: Theoretical Background. *International Journal of Engineering Research and Applications*, *3*(5), pp.605-610.

Jagielska, I. and Jaworski, J., 1996. Neural Network for Predicting the Performance of Credit Card Accounts. *Computational Economics*, *9*(1), pp.77-82.

Jo, H., Han, I. and Lee, H., 1997. Bankruptcy Prediction using Case-based Reasoning, Neural Networks, and Discriminant Analysis. *Expert Systems with Applications*, *13*(2), pp.97-108.

Kalimipalli, M. and Susmel, R., 2004. Regime-switching stochastic volatility and short-term interest rates. *Journal of Empirical Finance*, *11*(3), pp.309-329.

Kamruzzaman, J. and Sarker, R.A., 2003, Forecasting of Currency Exchange Rates using ANN: A Case Study. In *Proceedings of the 2003 IEEE International Conference on Neural Networks and Signal Processing,* 1, pp. 793-797.

Kim, S.H. and Chun, S.H., 1998. Graded Forecasting using an Array of Bipolar Predictions: Application of Probabilistic Neural Networks to a Stock Market Index. *International Journal of Forecasting*, *14*(3), pp.323-337.

Kiviluoto, K., 1998. Predicting Bankruptcies with the Self-organizing Map. *Neurocomputing*, *21*(1), pp.191-201.

Kokic, P., 2002. *A Multi-Layer Perceptron for Imputing Missing Values in Financial Panel/Time Series Data1*. Working Paper 5. QANTARIS GmbH, Frankfurt am Main.

Krishnan, T. and McLachlan, G., 1997. The EM Algorithm and Extensions. *Wiley*, *1*(997), pp.58-60.

Kyle, A.S., 1985. Continuous Auctions and Insider Trading. *Econometrica: Journal of the Econometric Society*, pp.1315-1335.

Lafferty, J., McCallum, A. and Pereira, F., 2001, Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML, 1*, pp. 282-289).

Lee, C. and Ready, M.J., 1991. Inferring Trade Direction from Intraday Data. *The Journal of Finance*, *46*(2), pp.733-746.

Leigh, D., 1995. Neural Networks for Credit Scoring. *Intelligent Systems for Finance & Business*, pp.61-69.

Leshno, M. and Spector, Y., 1996. Neural Network Prediction Analysis: The Bankruptcy Case. *Neurocomputing*, *10*(2), pp.125-147.

Linnainmaa, J.T. and Saar, G., 2012. Lack of Anonymity and the Inference from Order Flow. *Review of Financial Studies*, *25*(5), pp.1414-1456.

Little, R.J. and Rubin, D.B., 2014. Statistical Analysis with Missing Data. *John Wiley & Sons.*

Neal, R.M., 2012. Bayesian Learning for Neural Networks. *Springer Science & Business Media.*

Odders-White, E.R., 2000. On the Occurrence and Consequences of Inaccurate Trade Classification. *Journal of Financial Markets*, *3*(3), pp.259-286.

Olmeda, I. and Fernández, E., 1997. Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction. *Computational Economics*, *10*(4), pp.317-335.

Piramuthu, S., Shaw, M.J. and Gentry, J.A., 1994. A Classification Approach using Multi-layered Neural Networks. *Decision Support Systems*, *11*(5), pp.509-525.

Refenes, A.P. and Holt, W.T., 2001. Forecasting Volatility with Neural Regression: A Contribution to Model Adequacy. *IEEE Transactions on Neural Networks*, *12*(4), pp.850-864.

RJa, L. and Rubin, D.B., 1987. Statistical Analysis with Missing data. *John Wiley & Sons.*

Roderick, J., Little, A. and Rubin, D.B., 2002. Statistical Analysis with Missing Data. *J. Wiley*.

Rubin, D.B., 1976. Inference and Missing data. *Biometrika*, *63*(3), pp.581-592.

Rubin, D.B., 1996. Multiple Imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), pp.473-489.

Schafer, J.L. and Graham, J.W., 2002. Missing Data: Our View of the State of the Art. *Psychological Methods*, *7*(2), p.147.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. *CRC Press.*

Schafer, J.L., 1999. Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, *8*(1), pp.3-15.

Shleifer, A. and Vishny, R.W., 1997. The Limits of Arbitrage. *The Journal of Finance*, *52*(1), pp.35-55.

Smola, A.J. and Schölkopf, B., 2004. A Tutorial on Support Vector Regression. *Statistics and Computing*, *14*(3), pp.199-222.

Specht, D.F., 1991. A General Regression Neural Network. *IEEE Transactions on Neural Networks*, *2*(6), pp.568-576.

Sprott, D.A., 2008. Statistical Inference in Science. *Springer.*

Steiner, M. and Wittkemper, H.G., 1997. Portfolio Optimization with a Neural Network Implementation of the Coherent Market Hypothesis. *European Journal of Operational Research*, *100*(1), pp.27-40.

Tam, K.Y. and Kiang, M.Y., 1992. Managerial Applications of Neural Networks: The Case of Bank Failure Predictions. *Management Science*, *38*(7), pp.926-947.

Taskar, B., Abbeel, P. and Koller, D., 2002. Discriminative Probabilistic Models for Relational Data. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence* (pp. 485-492).

Teixeira, J.C. and Rodrigues, A.J., 1997. An Applied Study on Recursive Estimation Methods, Neural Networks and Forecasting. *European Journal of Operational Research*, *101*(2), pp.406-417.

Torsun, I.S., 1996. A Neural Network for a Loan Application Scoring System. *The New Review of Applied Expert Systems*, *2*, pp.47-62.

Warga, A., 1992. Bond returns, liquidity, and missing data. *Journal of Financial and Quantitative Analysis*, *27*(04), pp.605-617.

Williams, C.K. and Rasmussen, C.E., 2006. Gaussian Processes for Machine Learning. *The MIT Press*.

Yao-Hua Tan, W.T., 2000. Toward a Generic Model of Trust for Electronic Commerce. *International Journal of Electronic Commerce*, *5*(2), pp.61-74.

Yee, L.C. and Wei, Y.C., 2012. Current Modeling Methods used in QSAR/QSPR. *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, *2*, pp.1-31.

Yeo, A.C., Smith, K.A., Willis, R.J. and Brooks, M., 2002. A Mathematical Programming Approach to Optimise Insurance Premium Pricing within a Data Mining Framework. *Journal of the Operational research Society*, *53*(11), pp.1197-1203.

Yoon, Y., Guimaraes, T. and Swales, G., 1994. Integrating Artificial Neural Networks with Rule-based Expert Systems. *Decision Support Systems*, *11*(5), pp.497-507.

Zhu, H., 2006. An empirical comparison of credit spreads between the bond market and the credit default swap market. *Journal of Financial Services Research*, *29*(3), pp.211-235.