CAPTURING THE ORDER IMBALANCE WITH HIDDEN MARKOV MODEL: A CASE OF SET50 AND KOSPI50

Po-Lin Wu, **Wasin Siwasarit, Ph.D.**

ABSTRACT

Based on the empirical evidence of the recent strand of the literature, Market Efficiency creation process is not instantaneous, but rather attains over short-horizon of time. With the low liquidity market, the price movement of financial assets can be predicted by order imbalance indicators. In contrast, in a more liquidity market, the predictability of return is significantly decreased. In this study, we implement one of the well-known machine learning models for pattern recognition known as the Hidden Markov Model (HMM) with order imbalance to forecast the price movement of selected stocks in markets with different levels of liquidity which are the Stock Exchange of Thailand (SET) and Korea Exchange (KRX). As the consequence, we can create an algorithmic trading strategy based on the states of risky assets captured by the models. The result is consistent with the previous literature that both the predictability of the models and the profitability of the strategy diminish as the frequency decreases and market liquidity increases. Remarkably, our model in the market with lower liquidity is able to generate signal that achieves average hit ratio of 83.38% in predicting the risky assets' positive price movement at frequency of 5 minutes.

Keywords: Algorithmic trading, HMM, market efficiency, liquidity, order imbalance

Mr. Po-Lin Wu
Master of Science Program in Finance (International Program)
Thammasat University
Mail: bolinw@gmail.com
Mailing Address: 99 Moo 1, Nikompattana, Rayong Thailand 21180
Phone Number: +66-8-53910959

Wasin Siwasarit, Ph.D.
Faculty of Economics
Thammasat University
Mail: wasin@econ.tu.ac.th

INTRODUCTION

The Efficient Market Hypothesis defined by Fama (1970) states that the asset price should fully reflect all available information; thus, the asset return is not predictable and passive trading is always the optimal trading strategy. However, the empirical evidence from the study by Chordia T. and Roll R. (2005) based on 150 stocks listed on NYSE during year 1996, 1999 and 2002 shows that the market is not strong-form efficient; the future return of selected assets was predictable over the interval of 5 to 30 minutes by using the order imbalance.

The efficiency creating process is also affect by the market liquidity. The previous literature also shows that the predictability of order imbalance is linked to the liquidity of market. The research by Chordia T. et al (2008) provided strong evidence that at a more liquid regime, the predictability of asset tends to disappear due to investors taking advantage of low bid-ask spread.

Based on previous literatures, to beat the market, we should focus on the intra-day frequency, in which the market efficiency is possibly not attained. In addition to the literature of market efficiency, the quantitative hedge fund firms, such as Renaissance Technologies and Two Sigma in the US, were able to outperform the market by utilizing systematic and algorithmic trading and have been actively hiring professionals from field of information theory, which is a field that specializes in symbol and pattern recognition. Their success shows that, even in a market that is highly liquid, the market is still predictable at very high frequency. Therefore, the technology or the models that are utilized in the quantitative trading should be further studied.

This study aims to introduce the Hidden Markov Model and test its prediction ability in forecasting intra-days price movement of selected stocks in the SET50 index and the KOSPI 50 Index. We address the empirical evidence of return predictability by building a trading strategy and back-tested with the inclusion of transaction cost based on the patterns we discover with the proposed model. We also compare the performance of our model with the conventional buy and hold strategy in two different markets.

This paper contribute to 1) the advancement of algorithmic trading in Thailand 2) formulation of trading strategies for institutional or individual traders 3) study of the applicability of machine learning model in the Thai and Korean capital market 4) study of market efficiency in countries with different level of stock market liquidity at intraday frequency.

The paper is organized as follows. The next chapter documents review on literature. Chapter 3 describes the related theoretical framework of this study. Chapter 4 presents the methodologies of this study. Chapter 5 reports the result on both predictability and profitability of the HMM model. Last chapter contains the discussion and further recommendation of this study.

REVIEW OF LITERATURE

Hidden Markov Model in Financial Time Series Forecasting

The usage of Hidden Markov Model in Financial time series can be traced to a decade ago. Hassan, R. (2005) proposed a Hidden Markov Model with continuous emission to forecast the next day stock closing price of 4 different airline stocks. The model he proposed applies the intra-day high, low, open and closing price of stock to predict the next day closing price. However, the result was similar to performance of Artificial Neural Network and was unreliable in practical use. In order to improve the performance and accuracy of price prediction, Hassan, R. (2009) combined the Hidden Markov Based prediction method with fuzzy model to improve the accuracy of the model. The study was estimated with the same data set from the previous research for both training and testing. The result showed an improvement in prediction error in comparison to original HMM based prediction model, Artificial Neural Network and ARIMA. In the latest iteration, Hassan, R. (Hassan R. , 2013) improved the system by introducing the Adaptive Fuzzy Interference System which allowed the system to be able to adapt to the new arrival of data. The author applied the new system with 5 consecutive weekly stock index price data vectors to predict the weekly index movement and the

result showed improvement in accuracy over the previously proposed HMM-Fuzzy Model.

There were other researches proposed improvement or other approaches in training the Hidden Markov Model. Satish & Jerry (2010) compares the performance of prediction of Hidden Markov Model with Support Vector Machine in predicting the closing price of stocks in the S&P 500 Index. The paper proposed the k-mean algorithm for parameters initialization of Hidden Markov Model. The importance of initialization can be observed from the result; the hit rate of stock prediction decreased substantially. The initialization problem of Hidden Markov Model was not addressed in system proposed by Hassan, R. (2013), and thus further effort in investigating parameters initialization might be crucial to the prediction power of the model. Another research by Patrik, I. (2008) tried to build algorithmic trading strategy by applying both discrete and continuous Hidden Markov Model to predict the exchange rate of EURUSD. The author was able to make positive cumulative profit and obtain the Sharpe ratio 0.91 during the simulation period. Other than the exchange rate, the author also attempted to include other factors into the model. However, the result showed that the additional factors did not improve the model and worsen the profitability.

Order Imbalances and Stock Return

Volume is often used as a proxy in literature to describe the relationship between trading activity and market return. However, the order imbalance bears more information in term of trader's intent and direction of the stock price is headed.

The empirical evidence from the earlier research on the relationship between individual return and order imbalances by Chordia T, and Subrahmanyam A (2002) based on the daily NYSE data indicates that traders tend to split orders over period to mitigate price impact, which causes autocorrelated price pressure and results in a predictable relation between the imbalance and equilibrium price changes. The later research by Chordia T. and Roll R. (2005) also revealed that the future stock return can be predicted by the lagged order imbalance over the interval from 5 to 60 minutes; this evidence supports that the market is not efficient in the strong form. The further research by

Chordia T. and Roll R. (2008) on stock return, order flow and market liquidity highlighted the stylized fact that the predictability of individual stock return tends to disappear when the market is in a more liquid regime.

THEORETICAL FRAMEWORK

Efficient Market Hypothesis

The conventional investment theory proposed by Fama (1970) defined the efficient market as a market in which the prices always reflected the available information in three different considerations. In short summary, for weak form efficiency, the assets prices fully reflect the historical price; for the semi-strong form efficiency, the assets prices reflect all information that is publicly available, and the strong form efficiency, the assets prices reflect privileged information that is available to only specific participants. Consequentially, the result of such remark is that, in an efficient market, the prices of risky assets should accurately reflect the fundamental value, and thus no excess return can be generated from trading.

The empirical evidence over daily horizon seems to support the efficient market hypothesis; the previous literature by Chordia et al. (2005) showed that S&P Index follows random work and had insignificant auto-correlations despite the fact that public unavailable information was incorporated.

Market Efficiency, order imbalance and market liquidity

In an early research of market order imbalances on the S&P 500 by Chordia et al (2002) documented an interesting phenomenon; the market order imbalances (defined as daily aggregated purchase order less sell order) are highly predictable on the daily basis. Empirically, a day with high order imbalance will likely be followed by high order imbalance on the same side. However, given the predictability, the S&P 500 follows random work over a horizon of a day and had no auto-correlation at first or other longer lags. The observation implies that, some investors were able to correctly forecast the price pressure created by the order imbalances and exploit the price pressure, in which the trades are able to remove the auto-correlation of return within the horizon of one day.

Such phenomenon raises the question of how quickly the predictability of return is removed by the countervailing trades conducted by the investors who observed the order imbalance. However, it is certain that the process of removal of predictability of return is not instantaneous; it must take at least some time for investors to realize the information of order imbalances.

The further research by Chordia et al (2005) investigated the time taken for traders to take countervailing position that removes the predictability of returns. The result reconcile the belief that traders though do not have the information of order imbalance, but become aware of the information and take the countervailing position. Under the horizon of 30 minutes, the return is no longer predictable by using the order imbalances.

The empirical evidence also indicated that the speed of convergence is affected by the market liquidity. Chordia et al (2008) investigated the predictability of return using order imbalance in different liquidity regime. The result supported the evidence that the market in a more liquid regime is less predictable and is close to random walk. This observation implied in a liquid regime, information is more effectively incorporated into the price of risky assets. One rationale to explain this phenomenon is that due the smaller bid-ask spread in the liquid regime, informed traders have more incentive to submit the countervailing orders and thus catalyzed the speed of convergence.

METHODOLOGY FRAMEWORK

Order Imbalance

Based on the previous literature, the stock price movement can be predicted by the order imbalance indicator over a very short horizon. The research by Chordia et al (2005) defined the order imbalance in 3 different forms: the number of buy order less the number of sell order (OIB#), the number of buy-initiated shares purchased less the number of seller-initiated shares sold (OIBSh) and the dollars paid by buy-initiators less the dollars received by sell-initiators (OIB$). The last two factors OIBSh and OIB$ have empirically better predictability of future return in comparison to OIB#, but all three

informations are only available to market makers or traders who are able to estimate the imbalance in the New York Stock Exchange.

For the target markets of this study, we have both sell order and buy order data widely available to the public, and hence the order imbalance indicator will be constructed based on the available information. We approximate our order imbalance indicator in a similar approach to the recent research by Shen D. (2015) known as the Order Imbalance Ratio (OIR).

The OIR measures the size of buy order in relative to the sum of number of buy orders and number of sell orders at a specific time point. Thus, the low value of order imbalance ratio implies that there is lack of demand or excess of supply on a particular asset; whereas, the high value of order imbalance ratio implies that there is excess demand or lack of supply on a particular asset.

The order imbalance will be expressed as a relative term. The reason behind using this method to construct the indicator is that we can quantize the indicator with ease and scale down the indicator.

$$OIR_t = \frac{V_t^B}{V_t^A + V_t^B}$$

$$Where\ V_t^B = volume\ \ of\ current\ best\ bid\ price$$

$$V_t^A = volume\ of\ curret\ best\ ask\ price$$

Data Quantization

For discrete case of the Hidden Markov Model, the discretization process needs to be conducted to convert both return and order imbalance indicator into representative symbols.

We classify the price movement into two categories; the price moves down and price remains the same or price moves up. On the other hand, there is no clear guideline on discretizing the order imbalance ratio denoted OIR; therefore, we separate the order imbalance into 3 groups, which are the groups with OIR in the 25 percentile, OIR above the 75 percentile and the OIR that is between 25 and 75 percentile. The 25 percentile and 75 percentile are approximated by averaging the 25 percentile and 75 percentile of each

stock at each frequency during the pre-study period. Table 1 reports the detail of data quantization for the discrete Hidden Markov Model.

Table 1: Data Quantization for discrete the Hidden Markov Model

Symbol	Return interval	Order Imbalance Ratio
1	$0\% <$	$< 25\%\ percentile$
2	$0\% <$	$25\%\ percentile \leq OIR \leq 0.75\%\ percentile$
3	$0\% <$	$OIR > 75\%\ percentile$
4	$\geq 0\%$	$< 25\%\ percentile$
5	$\geq 0\%$	$25\%\ percentile \leq OIR \leq 0.75\%\ percentile$
6	$\geq 0\%$	$OIR > 75\%\ percentile$

Noted: For symbol 1, 2 and 3, the return interval can be interpreted as negative price movement, i.e. $\Delta P < 0$; whereas, the return interval of symbols 4, 5 and 6 can be interpreted as price movement that is not negative, i.e. $\Delta P \geq 0$.

Table 2: 25% and 75% percentile of Order Imbalance Ratio

Market	Frequency	25% Percentile	75% Percentile
SET50	5 minute	0.40	0.65
	10 minute	0.40	0.64
	30 minute	0.41	0.62
KOSPI50	5 minute	0.34	0.61
	10 minute	0.35	0.60
	30 minute	0.38	0.59

Noted: The percentiles are computed based on the data from 1st October 2016 to 31st October 2016. The percentiles are computed for each individual stock then are averaged to obtain the value in the table.

Hidden Markov Model

The Hidden Markov is a statistical model that is designed to capture the dynamic that cannot be directly observed from a set of observations. The simple discrete Hidden Markov Model mainly consists of two parts, first a set of unobservable states $S = \{s_1, s_2, \dots, s_T\}$ and a set of observable symbol $O = \{o_1, o_2, \dots, o_T\}$. At each step/time slot t, the movement of state to another state is governed by a set of transition probability. The

sequence of observable symbols is a state dependent process, i.e. each state governs a probability distribution of observable symbols.

The reason of applying Hidden Markov Model in this study is because of its ability to capture the hidden dynamic or behavior of stock market. In this study, we aim to capture the hidden state of order imbalances through the observable symbols of stock price movement and buy/sell order movement in a confident manner. The state of order imbalance can be interpreted as a state where new information has not yet adjusted into the asset price or the state where the asset price deviated from the fundamental. If the model is able to capture the order imbalance state in a consistent and confident manner, then it is possible to profit from the price pressure created by the order imbalance state.

Three Fundamental Problems of Hidden Markov Model

The characterization of a Hidden Markov Model can be described as following: 1) Number of states in the Model 2) Number of observable symbols in the model 3) the probabilities of state transition 4) the emission probability distribution of observable symbols generated from states 5) the prior probability distribution of initial states. For the rest of paper, following notations will be used.

N = Number of states in the model

M = Number of observable symbols

T = Length of observable symbols sequence

H = A set of possible states in the model, H = $\{h_1, h_2, \dots, h_N\}$

O = The observable symbols sequence, $O = \{o_1, o_2, \dots, o_T\}$

S = The states sequence, $S = \{s_1, s_2, \dots, s_T\}$

A = The $N \times N$ state transition matrix

B = The $N \times M$ observable symbols emission matrix

α_{ij}= The probability of transition from state i to state j

$b_j(o_t) =$ The probability of generating observation t at state j

Π = The $1 \times N$ vector of prior probability of each state

$\pi_i =$ Initial probability of starting in state i

$\boldsymbol{\lambda}$ = The Hidden Markov Model, consisted of A, B and π. $\lambda = (A, B, \boldsymbol{\pi})$

The three fundamental problems of a Hidden Markov Model are of the following:

1. The Evaluation Problem: Given the model $\boldsymbol{\lambda}$, Compute the probability of the observed sequence of symbols i.e. compute $P(O|\boldsymbol{\lambda})$.
2. The Decoding Problem: Given both the model $\boldsymbol{\lambda}$ and the observed sequence of symbols, what is the most likely state sequence?
3. The Learning Problem: Given observation sequence and possible parameters of model i.e. A, B and Π, adjust the parameters to find the model that best explain the observed sequence, i.e. find $\boldsymbol{\lambda}$ that maximizes $P(O|\boldsymbol{\lambda})$.

The evaluation problem is used in the learning problem to test for convergence to the local maxima. The forward or backward algorithms are used to solve this problem.

The decoding problem finds the most likely state sequence given the model and observation sequence. In this study, the Viterbi algorithm will be applied to solve the problem; it is an algorithm that finds the state sequence of a fixed observation sequence with the maximum likelihood i.e. $argmax\ P(S, O|\boldsymbol{\lambda})$.

Last but not least, the learning problem of Hidden Markov Model finds the model parameters $\boldsymbol{\lambda}$ that best explain the observed sequence (maximizing the probability $P(O|\lambda)$). The learning problem cannot be solved analytically, and is conventionally solved by the Expectation Maximization algorithm called the Baum-Welch algorithm.

Number of states in the Hidden Markov Model

As discussed in the literature review section, the number of states in the Hidden Markov Model can be interpreted as different behavior of markets. Determining the optimal number of states in the market would be crucial to the trading signal generation of the model. The number should not be too large, there is little to no distinction between each state; on the other hand, if the number of states also should not be too small, then the model may not be able to capture the hidden behaviors of market movement. For this study, we set the minimum number of states of stock to three, in which the three states represent the information of asset price being overvalued, undervalued or in the

equilibrium. However, there are possibly unknown hidden states in the market; the model might be improved if we increase the number of states for coverage of other hidden states. For the scope of this study, we aim to test the performance of our model from 3 states to 5 states.

Generating Trading Signals

Discrete Case

From the discussion in the literature review section, we can use the solution to the learning problem to approximate the best model for the given observation sequence. By using the model, from the decoding problem we can find the probability of each state that generate the current observation and find the most probable state that generates the current symbol.

By knowing the most probable state at t, we can utilize the transition matrix A estimated in the learning problem to find out the likely transition and predict the state at $t+1$. Then, based on the predicted state, we determine the probability of observing certain asset price movement by using the emission matrix B. In order to be more certain about the outcome in the next time period, a threshold needs to be imposed. For the purpose of this study, we are interested in observing a state that has confident transition and follow by a state where the probability of upward price movement is high. Therefore, based on table 4.3, the threshold can be set as the probability of observing symbol 4, 5 and 6 at $t+1$. The value of certainty of outcome at $t+1$ can be determined as follow.

$$P\left(s_{t+1} = h_i \middle| s_t = h_j\right) \cdot b_i(o_t)$$

If the p-value of is greater than defined threshold, then the trading signal of entering position is generated, otherwise we liquidate current position. In this study, we aim to capture the order imbalance state in a consistent and confident manner; thus, it would be in our interest to filter out the signals with lower confidence level to avoid excess loss from transaction cost and incorrect predictions. To set up our threshold, we propose a 90% confidence level in transition and 90% confidence level in observing positive or no price movement. The joining two confidence level, we propose to set the threshold at value of 80%.

Continuous Case

For the continuous case of Hidden Markov Model, each hidden state is associated with the probability density function of observables instead of discretized probability for each possible observable. Therefore, the trading signals are generated based on the interpretation on the properties of probability distribution function.

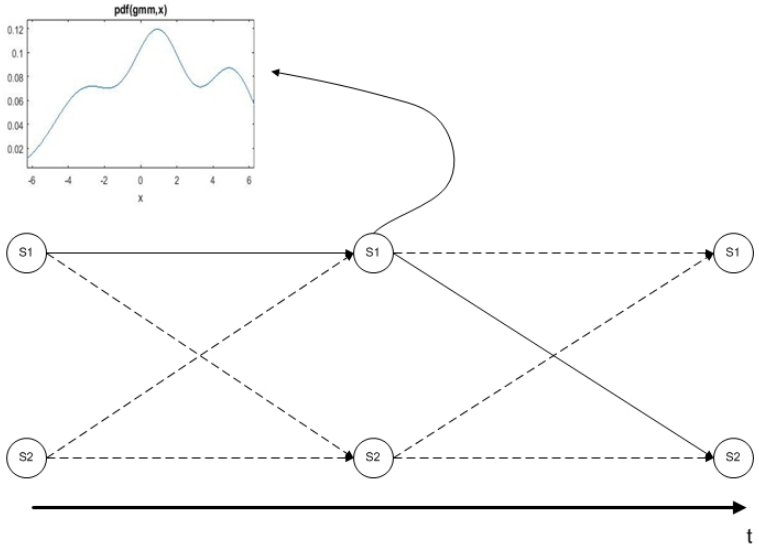
The result on normality test indicates that the probability distribution of return and OIR are not normal at 95% confidence level. With support of empirical evidence that the probability distributions of observables are not normal, we turn our attention to the Gaussian Mixture Model to better describe the properties of probability distribution of observables.

HMM with the Gaussian Mixtures

As discussed in the previous sub sections, the observables are continuous. Therefore, instead of the emission matrix of observables, we have the parameters for the probability density function of Gaussian Mixture Model as shown in figure 1. As a result, the probability of observables generated from a particular state at time t defines as:

$$b_j(o_t) = \sum_{i=1}^{M} w_i P(o_t | G_j\,(O|\mu_i, \Sigma_i))$$

Figure 1: 2 states HMM with Gaussian Mixture Model of 3 components



(13)

In this study, we propose 2 approaches to generate trading signal and are discussed below:

Approach I: Apply only first moment:

To generate trading signal, we first set-up a threshold level of return 0**.** At each trading interval, the Viterbi algorithm is first used to determine the most probable state at current t. Then, we utilized the trained transition matrix to determine the state at $t+1$. The expected return is then calculated by using the mean return of each Gaussian component. The equation is defined as follow:

$$E[r] = \sum_{i=1}^{M} w_i \mu_{i,return}$$

If the expected return is greater than 0, then trading signal is generated; whereas, if the expected return is less than 0, then we liquidate the position.

Approach II: Apply both first moment and second moment

Based on this approach to generate the trading signal, we first set-up 2 thresholds, a threshold of confidence level of next period return 0.8 and a threshold for expected return 0**.** At each trading interval, the Viterbi algorithm is firstly used to determine the most probable state at current time t. Then, we utilized the trained transition matrix to determine the future state at $t+1$. The p-value is then calculated by using the following equation:

$$p-value = P\left(s_{t+1} = h_i \middle| s_t = h_j\right) \cdot \sum_{i=1}^{M} w_i \, P(return \geq 0 | G(X|\mu_i, \Sigma_i))$$

If calculated p-value is greater than or equal to the defined threshold, then the trading signal is then generated.

Trading Strategies

In the previous section, the paper has discussed about how the signal is generated from the Hidden Markov Model. Due to the nature of discretization method, the generated signal can only predict the direction of movement, but not the size movement. Thus the strategy is a form of gambling with the belief that there will be more gains than

losses from the gambling. In this study, we propose a simple algorithm to handle the signals generated from the Hidden Markov Models:

1. Train the Hidden Markov Model for each stock
2. Obtain a list of stocks (trading signals) that we should enter long position.
3. Liquidate all stocks that are current in long position and are not in the list.
4. If there is any remaining wealth, allocate the wealth equally to all stocks in the list.
5. If at the end of the day, then go to step 1, else go to step 2
6. Iterate until the end of observations

The mid-point closing price at the end of each interval will be used as the trading price for buying and selling the shares and bi-directional transaction cost at level of 0.05% is used to assess applicability of the model for different group of investors.

Performance Measurement

Benchmark

The SET and KOSPI Total Return Index will be used as benchmark to compare with the profitability of the trading strategy. The Total returns index is calculated based on the assumption that all dividends are immediately re-invested.

Hit Ratio

The hit ratio will be used to assess the performance of Hidden Markov Model on forecasting the stock price movement of out-of-sample data set. The calculation of Hit Ratio will be separated from the trading strategy and will be calculated for each individual stock.

The hit ratio is defined as follow:

$$Hit\ Ratio = \frac{h}{n}$$

Where n = total number of trading signals that results in positive/negative price movement

h = total number of trading signals that correctly predict the positive price movement

To test whether or not the forecast is more than just coin flip guess, we will apply one sample t-test to test the null hypothesis whether or not the hit ratio is equal to 0.5. The formula of one sample t-test is defined as follow:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Where $\overline{x}$ = sample mean

$\frac{s}{\sqrt{n}}$ = standard error

We will use one-sample t-test to test the hypothesis of $H_0: Hit\ Ratio = 0.5$, $H_a: Hit\ Ratio \neq 0.5$. Under Efficient Market Hypothesis, the null hypothesis should not be rejected; if the null hypothesis is rejected, then there exists pattern in the stock market and the price movement can be predicted.

Jensen's Alpha

Alpha is the abnormal rate of return that exceeds the expected return at given risk defined by a specific model. Under the efficient market hypothesis, the alpha should be insignificant and equal to 0, because it is not possible to outperform the overall market. For this study, the Capital Asset Pricing Model (CAPM) will be used to estimate whether there exist significant alpha for our back-testing portfolio over the horizon of three months. The model is defined as:

$$E[r_i] - r_f = \alpha_J + \beta(E[r_M] - R_f)$$

Where r_i = portfolio return

r_f = risk free rate

r_M = market return

Data

We conduct this study in markets with different level of liquidity: the Thai stock market and Korean stock market. In particular, the Korean market is more liquid than the Thai capital market. As reported by the World Bank, in year 2015, the annualized stock turnover ratio of Stock exchange of Thailand (SET) is roughly 77.8%; in contrast, the annualized stock turnover ratio of Korea Exchange (KRX) is roughly 149.8%, which is more than 2 times of turnover ratio of Thai stock market. Based on the theory, we expect

that the Korea Exchange should have lower predictability of asset return due having higher liquidity; in other word, our model should perform relatively poor in Korea stock market in comparison to Thai stock market.

We limit the scope of this study to stocks listed in SET50 Index and KOSPI50 Index; the SET50 Index is chosen because the stocks are relatively more liquid in comparison to other stocks listed in the SET and are the stocks with large market capitalization, thus mitigate the issues of no trades. To compare between markets, we decide to pick KOSPI50 index in Korea that selects the stocks in similar method of SET50. To further limit the scope of this study, we reduce the number of stocks to 10 for each market and the selection method is described below.

The stocks in this study are selected by going through the following steps

1. Filtered stocks that are not consistently listed in SET50 Index and KOSPI 50 Index during the period form 1st January 2012 to 31st July 2016
2. Keep the top 10 stocks with highest average volume turnover in the respective market to ensure liquidity of stocks. The turnover is calculated by using the following formula**:**

$$Volume\ Turnover = \frac{250\ days\ average\ daily\ volume\ turnover\ (as\ of\ 21st\ November\ 2016)}{Average\ Total\ Common\ shares\ outstanding(Through\ out\ year\ 2015)}$$

The list of stocks applied in this study is represented in the table 3 and 4.

Table 3: The listed stocks selected from SET50 for study

Ticker	Company Name	Sector
ADVANC.BK	Advance Info Service PCL	Information & Communication
BANPU.BK	Banpu PCL	Energy & Utilities
BCP.BK	Bangchak Petroleum PCL	Energy & Utilities
CPF.BK	Charoen Pokphand Foods PCL	Food and Beverage
DTAC.BK	Total Access Communication PCL	Information & Communication
IRPC.BK	IRPC PCL	Energy & Utilities
IVL.BK	Indorama Ventures PCL	Petrochemicals & Chemicals
PTTEP.BK	PTT Exploration and Production PCL	Energy & Utilities
TCAP.BK	Thanachart Capital PCL	Banking
TRUE.BK	True Corporation PCL	Information & Communication

Source: Author's calculation

Table 4: The listed stocks selected from KOSPI 50 for study

TICKER	Company Name	Sector
034220.KS	LG Display Co, Ltd	Electrical & Electronic Equipment
066570.KS	LG Electronics Inc	Electrical & Electronic Equipment
051910.KS	LG Chem Co, Ltd	Chemicals
005490.KS	POSCO	Iron & Metal Products
006400.KS	Samsung SDI Co, Ltd	Electrical & Electronic Equipment
009150.KS	Samsung Electro Mechanics Co Ltd	Electrical & Electronic Equipment
010140.KS	Samsung Heavy Industry Co, Ltd	Transport Equipment
000880.KS	Hanwha Corp	Finance
000720.KS	Hyundai Engineering & Construction Co Ltd	Construction
009540.KS	Hyundai Heavy Industry Co, Ltd	Transport Equipment

Source: Author's calculation

As discussed in the previous section, the input of the model are bid size, ask size, closing price at interval of 5, 10 and 30 minutes. Based on these data, order imbalance indicator and return are computed. The model uses rolling window technique and the window will be move by 1 day for every step; the size of window are 15 trading days for 5 minute data, 19 trading days for 10 minute and 40 trading days for 30 minutes data. The training is conducted on the daily basis; this means that the model will be re-trained when the market closed. The actual size of rolling window is defined by:

$$Size = Number\ of\ interval\ per\ day \times number\ of\ days$$

The data will be divided into 2 periods: the pre-study period and the back-testing period. The data in the pre-study period will be used for both initial training of model and estimation of 25 and 75 percentile of order imbalance ratio. The trading period will be the period for test both profitability and accuracy in this study. The pre-study period starts from 1st October 2016 to 31st October 2016; the study period begins from 1st November 2016 and ends in 31st January 2017, with the exception on 30 minutes case due to wider range of available data.

For performance comparison, we collected the 3-month Bangkok Interbank Offered Rate (BIBOR) and Korea Interbank Offered Rate (KORIBOR) as the proxy to risk-free rate. The daily and monthly total return index of Stock Exchange of Thailand and Korea Exchange (KRX) are collected as our benchmark.

The intra-day data of selected stocks are collected from Reuter Eikon, the daily and monthly 3-month BIBOR rate is collected from database of Bank of Thailand. The daily and monthly 3-month KORIBOR rate is collected from Korean Statistical Information Service.

RESULT AND DISCUSSION

Predictability of the Discrete HMM for selected stocks in SET50

Based on table 5, the models perform more confidently and consistently in the Thai market; our basic case, the basic 3 states discrete Hidden Markov Model is able to achieve hit ratio of average 78.61% at 5 minute frequency. The model seems to improve at 5 minute frequency as we increase the number of states; at 5 states, the average hit ratio increases to 83.38% with no predictability lower than 70% for each individual risky asset.

As we lower the frequency, the hit ratio decreases and the models become less confident in making a prediction. Moving from 5 minute frequency to 10 minute frequency, the average hit ratio lowers to 70.59% and at frequency of 30 minutes, the average hit ratio of discrete models decreases to 65.63%.

The result seems to support the hypothesis that the order imbalance tends to lose its predictability as the interval enlarge; our model becomes less confident in making predictions (in the form of generating less signals) with cases that no prediction was made. This shows that, as the time increases, it becomes more difficult for the Hidden Markov Model to recognize a pattern. This evidence is consistent with the previous literature by Chordia et al (2005) that the information is adjusted into the price as time increases, thus the individual stock price becomes less predictable and follows random walk.

Predictability of the Discrete HMM for selected stocks in KOSPI50

In table 6, we report the result of our models in the Korean stock market produces evidence that the market liquidity does enhance the speed of convergence to efficiency. Compared to the model performance in Thai market, the models though were able to

achieve some predictability, seem to be much less confident in making predictions. Begin with the 5 minute frequency; the models generate relatively low number of signals in comparison to our model performance in the Thai market. Though the daily trading period in the Korean capital market is longer than the Thai market, the total number of signals generated is significantly less; over the horizon of 3 month, the average total number of signals generated for the Korean market is 289 signals. In addition, the 3 states and 4 states model were only able to generate signals for less than half of the selected stocks. At interval of 30 minutes, the average total number of signals decreased to 71 signals over the horizon of 3 months.

In term of hit ratio, the result indicates better models performance in lower frequency and tends to probability of coin toss in lower frequency. At 5 minute frequency, the average hit ratio of 3 states, 4 states and 5 states models is approximately 67.74% (see table 6), with the 5 states model performs in a more consistent manner (generated the most signals and achieved average hit ratio of 71.57%) . As the frequency decreases, the hit ratio indicates that the signals generated by models were no longer able to predict the price movement.

The result is consistent with the previous literature that the market liquidity enhances the speed of convergence to efficiency. Our models are significantly less confident in generating a signal and achieve low hit ratio in comparison to the same models performance in the Thai market. The finding seems to be consistent with the literature by Chordia et al (2008); when the market is in a more liquid regime, the bid-ask spread tends to narrower; as a result, the traders who observe the order imbalance will then have more incentives to take position and gain from the deviation of asset price from the fundamental. Such actions enhance the speed of adjustment of asset price, and our models become less confident in capturing the price pressure created by order imbalance due to the fact that the process of price adjustment to new information already occurred.

Predictability of the continuous model

As discussed the methodology section, we incorporate two approaches to generate signals for predicting the market movement. For the first approach, the signal is

generated by using the mean return of the predicted state; whereas, for the second approach, the signals are generated by calculating the probability of observing positive price movement. All in all, regardless which approach we use, our result of continuous model indicates that the models failed to capture the order imbalance states across all frequencies and number of states.

The first approach, i.e. predicting the movement by the mean return of state, also failed to generate a meaningful result. As shown in table 7 and 8, the hit ratios of the continuous model for all stocks, across all frequencies, are around the number of 0.5. Unsurprisingly, our t-test result indicates that the hit ratios for all cases of continuous models are not statistically deviated from 0.5, and we failed to reject the null hypothesis that the hit ratios are equal to 0.5. This result indicates that the continuous Hidden Markov Model failed to capture the order imbalance state, and the predictability is the same as a coin toss.

On the other hand, the approach II i.e. making prediction based on the probability distribution of return, the models were unable to generate p-value higher than 60% in both markets; therefore, the model was unable to generate a single signal due to our requirement of capturing the order imbalance state in a confident and consistent manner and the result was not recorded.

One possible explanation of why continuous models failed to produce meaningful result is the assumption of distribution. Due to the fact that intra-day return and order imbalance indicator are not normal, we attempt to mitigate the issue by assuming the three components Gaussian Mixture Models. However, the resulting Gaussian Mixture Model might not be enough to mitigate the extreme Kurtosis value of intra-day data. Comparatively, the BIC score of the 3 components Gaussian Mixture Model is better than the Gaussian Model, but not significantly better. As a result of excess kurtosis, the models suffer from the assumption and thus failed to produce meaningful result.

Table 5: Hit ratio of the Discrete HMM for selected stocks in SET50

The table shows the hit ratio of the Discrete Hidden Markov Model on each risky asset selected from SET50 Index with different number of states at different level of frequency. The hit ratio is calculated by number of generated signals that correctly predicted the future positive movement divided by number of signals generated that result in either negative or positive price movement. Any signals that resulted in zero mid-point price movement are ignored.

Frequency	# of states	ADVANC.BK	BANPU.BK	BCP.BK	CPF.BK	DTAC.BK	IRPC.BK	IVL.BK	PTTEP.BK	TCAP.BK	TRUE.BK
5 min	3	82.47%	66.92%	81.19%	64.88%	82.72%	78.64%	78.17%	74.10%	89.68%	87.33%
	4	77.43%	75.00%	89.86%	80.30%	83.85%	83.18%	74.38%	71.67%	86.47%	83.82%
	5	80.12%	80.70%	89.42%	85.31%	84.77%	89.19%	80.59%	70.73%	88.46%	84.46%
10 min	3	72.22%	73.68%	91.55%	85.37%	57.58%	72.86%	70.27%	54.17%	71.43%	72.97%
	4	86.67%	53.85%	66.67%	69.05%	66.67%	100.00%	63.64%	-	-	72.73%
	5	73.21%	44.68%	85.71%	77.22%	73.44%	85.29%	62.26%	40.00%	66.67%	62.96%
30 min	3	60.00%	67.86%	87.50%	69.44%	72.22%	72.97%	78.43%	50.00%	83.87%	75.76%
	4	40.00%	25.00%	100.00%	40.00%	80.00%	57.14%	65.52%	75.00%	-	83.33%
	5	70.00%	53.85%	60.00%	50.00%	83.33%	62.50%	68.75%	71.43%	44.44%	57.69%

Source: Author's calculation

Table 6: Hit ratio of the Discrete HMM for selected stocks in KOSPI50

The table shows the hit ratio of the Discrete Hidden Markov Model on each risky asset selected from KOSPI50 Index with different number of states at different level of frequency. The hit ratio is calculated by number of generated signals that correctly predicted the future positive movement divided by number of signals generated that result in either negative or positive price movement. Any signals that resulted in zero mid-point price movement are ignored.

Frequency	# of states	034220	066570	051910	005490	006400	009150	010140	000880	000720	009540
5 min	3	-	-	78.13%	72.62%	67.39%	-	-	-	-	-
	4	-	-	66.67%	80.00%	0.00%	52.63%	-	-	-	95.35%
	5	53.97%	76.92%	80.56%	77.14%	72.73%	61.73%	66.67%	-	75.00%	79.44%
10 min	3	58.33%	42.42%	70.69%	59.52%	75.00%	-	100.00%	50.00%	-	73.33%
	4	20.00%	62.07%	57.30%	57.14%	80.00%	100.00%	100.00%	50.00%	-	100.00%
	5	83.33%	50.00%	55.74%	40.00%	50.00%	100.00%	40.00%	-	42.86%	72.73%
30 min	3	-	-	-	60.00%	-	-	-	-	-	-
	4	-	0.00%	71.43%	48.72%	83.33%	50.00%	-	-	0.00%	47.83%
	5	-	25.00%	66.67%	50.00%	66.67%	83.33%	33.33%	42.86%	0.00%	37.50%

Source: Author's calculation

Table 7: Hit ratio of the Continuous HMM for selected stocks in SET50

The table shows the hit ratio of the Continuous Hidden Markov Model on each risky asset selected from SET50 Index with different number of states at different level of frequency. The prediction method is the mean return of the state predicted by the Viterbi algorithm. The hit ratio is calculated by number of generated signals that correctly predicted the future positive movement divided by number of signals generated that result in either negative or positive price movement.

Frequency	# of states	ADVANC.BK	BANPU.BK	BCP.BK	CPF.BK	DTAC.BK	IRPC.BK	IVL.BK	PTTEP.BK	TCAP.BK	TRUE.BK
5 min	3	50.46%	49.04%	49.88%	49.57%	50.81%	50.35%	50.68%	50.08%	50.87%	48.85%
	4	50.90%	48.14%	50.11%	49.15%	50.00%	50.00%	51.23%	50.19%	50.80%	48.53%
	5	50.55%	48.29%	49.83%	49.59%	50.12%	50.00%	50.56%	50.87%	50.70%	48.28%
10 min	3	51.68%	47.65%	50.36%	47.58%	49.50%	50.00%	50.80%	48.63%	50.11%	48.98%
	4	51.07%	47.88%	49.85%	49.70%	48.97%	49.46%	49.68%	49.22%	49.44%	48.03%
	5	50.41%	47.60%	50.34%	49.83%	48.52%	49.89%	49.90%	48.71%	50.63%	48.99%
30 min	3	51.24%	47.46%	50.48%	42.42%	53.16%	52.38%	51.48%	48.68%	52.28%	46.67%
	4	51.75%	48.15%	49.65%	48.67%	52.82%	51.00%	50.15%	49.11%	53.31%	46.34%
	5	49.57%	46.54%	50.58%	44.88%	52.50%	50.97%	51.54%	49.34%	52.69%	47.85%

Source: Author’s calculation

Table 8: Hit ratio of the Continuous HMM for selected stocks in KOSPI50

The table shows the hit ratio of the Continuous Hidden Markov Model on each risky asset selected from KOSPI50 Index with different number of states at different level of frequency. The prediction method is the mean return of the state predicted by the Viterbi algorithm. The hit ratio is calculated by number of generated signals that correctly predicted the future positive movement divided by number of signals generated that result in either negative or positive price movement.

Frequency	# of states	034220	066570	051910	005490	006400	009150	010140	000880	000720	009540
5 min	3	49.80%	51.42%	51.69%	50.13%	50.87%	49.79%	47.62%	48.80%	50.30%	49.17%
	4	50.24%	51.66%	50.59%	49.82%	50.86%	49.51%	48.56%	48.93%	51.55%	49.32%
	5	50.23%	51.63%	50.70%	50.10%	51.00%	49.48%	50.47%	48.55%	49.51%	48.74%
10 min	3	49.66%	51.43%	50.11%	49.22%	51.39%	50.11%	49.26%	48.54%	49.68%	49.13%
	4	49.70%	52.06%	50.66%	49.53%	51.47%	49.94%	48.24%	47.93%	47.46%	49.10%
	5	49.06%	51.98%	50.78%	49.62%	51.58%	47.41%	48.92%	49.39%	48.35%	49.05%
30 min	3	44.02%	53.54%	50.11%	48.80%	54.63%	48.08%	45.87%	49.63%	49.85%	47.95%
	4	46.86%	53.37%	49.88%	47.73%	54.05%	50.56%	47.74%	46.68%	49.50%	45.51%
	5	45.82%	53.87%	48.32%	47.98%	54.23%	50.14%	46.09%	46.63%	48.19%	45.33%

Source: Author's calculation

Profitability of the discrete models in both Thai and Korean stock market

As reported in table 10, with the assumption of 0.05% bi-directional cost, the result indicates that even at the highest frequency, our trading strategy was not able to achieve significant alpha in the Korean market. In contrast, as reported in table 9, the strategy shows a promising result of yielding significant positive alpha in all 5 minute cases and several cases at the 10 minute and 30 minute frequencies. The result shows that it is possible for institutional traders to make a profit by trading based on the predicted market movement.

However, it is to be noted that the result is under assumption of trading with mid-point price than the actual bid-ask price. Based on the previous literature, the reason of the strategy works in market with lower liquidity is because of lack of incentives for sophisticated traders to take position due to higher bid-ask spread.

Profitability of the continuous models in both Thai and Korean stock market

As discussed in section of predictability, the continuous models failed to capture the order imbalance state. Similarly, trading strategy using the continuous models failed to generate a significant and meaningful result.

The results of our regression analysis provided in table 11 and 12 indicate that there exists no significant alpha for all models, across all frequencies and in both market. This result is expected due to the fact that the continuous models show no predictability and fail to capture to order imbalance states because our assumption on distribution is unable to describe the properties of intra-day data. As a consequence, our models are lack of predictability and are unable to obtain a significant return from the strategy.

Table 9: Jenson's Alpha (Discrete HMM, SET50, 0.05% transaction cost)

The following tables compare the trading performance of the discrete Hidden Markov Model in SET50 over the horizon of 1st November 2016 to 31st January 2017. The independent variable is the excess daily return of portfolio and the dependent variable is the excess daily market return of SET. The adjusted daily 3-month Bangkok Interbank Offered Rate is used as a proxy to risk-free rate. For this set of data, institutional investor's level of transaction cost (0.05%) is assumed for estimation of return. All standard errors are Heteroskedasticity-robust standard errors.

Frequency	States		Coefficient	SE	t-stat	p-value
5 min	3 states	Intercept	0.02872	0.00538	5.33643	0.00000
		slope	4.40587	1.80102	2.44632	0.01738
	4 states	Intercept	0.03571	0.00494	7.23223	0.00000
		slope	3.93404	1.78874	2.19934	0.03172
	5 states	Intercept	0.04734	0.00399	11.86487	0.00000
		slope	1.68980	1.43897	1.17431	0.24491
10 min	3 states	Intercept	0.01056	0.00233	4.52421	0.00003
		slope	2.07871	0.91569	2.27009	0.02681
	4 states	Intercept	0.00145	0.00175	0.82488	0.41271
		slope	0.17092	0.43919	0.38916	0.69853
	5 states	Intercept	0.00409	0.00344	1.19164	0.23809
		slope	3.18249	1.40427	2.26630	0.02705
30 min	3 states	Intercept	0.00770	0.00157	4.89584	0.00001
		slope	0.33346	0.71038	0.46941	0.64048
	4 states	Intercept	0.00050	0.00137	0.36742	0.71460
		slope	-0.43771	0.78439	-0.55802	0.57890
	5 states	Intercept	0.00055	0.00134	0.41327	0.68088
		slope	0.76181	0.41203	1.84892	0.06940

Source: Author's calculation

Table 10: Jenson's Alpha (Discrete HMM, KOSPI50, 0.05% transaction cost)

The following tables compare the trading performance of the continuous Hidden Markov Model in KOSPI50 over the horizon of 1st November 2016 to 31st January 2017. The independent variable is the excess daily return of portfolio and the dependent variable is the excess daily market return of KOSPI. The adjusted daily 3-month Korea Interbank Offered Rate is used as a proxy to risk-free rate. For this set of data, institutional investor's level of transaction cost (0.05%) is assumed for estimation of return. All standard errors are Heteroskedasticity-robust standard errors.

Frequency	States		Coefficient	SE	t-stat	p-value
5 min	3 states	Intercept	-0.00085	0.00062	-1.36643	0.17682
		slope	0.12419	0.11424	1.08710	0.28127
	4 states	Intercept	0.00045	0.00083	0.54786	0.58579
		Slope	0.17233	0.08384	2.05555	0.04411
	5 states	Intercept	0.00028	0.00155	0.18078	0.85714
		Slope	0.35100	0.19263	1.82211	0.07334
10 min	3 states	Intercept	-0.00099	0.00102	-0.97218	0.33480
		Slope	-0.20259	0.17676	-1.14613	0.25622
	4 states	Intercept	-0.00276	0.00090	-3.07415	0.00316
		Slope	-0.01946	0.09214	-0.21116	0.83347
	5 states	Intercept	-0.00166	0.00078	-2.14127	0.03626
		slope	-0.04272	0.05371	-0.79545	0.42944
30 min	3 states	Intercept	-0.00042	0.00036	-1.16637	0.24800
		slope	-0.17758	0.10339	-1.71762	0.09094
	4 states	Intercept	-0.00083	0.00084	-0.99094	0.32563
		slope	-0.16460	0.14618	-1.12596	0.26459
	5 states	Intercept	-0.00172	0.00095	-1.81365	0.07465
		slope	0.02500	0.12511	0.19981	0.84230

Source: Author's calculation

Table 11: Jenson's Alpha (Continuous HMM, SET50, 0.05% transaction cost)
The following tables compare the trading performance of the continuous Hidden Markov Model in SET50 over the horizon of 1st November 2016 to 31st January 2017. The prediction method of mean return (please see method I of section 4.4.4.3) is used for generating signal. The independent variable is the excess daily return of portfolio and the dependent variable is the excess daily market return of SET. The adjusted daily 3-month Bangkok Interbank Offered Rate is used as a proxy to risk-free rate. For this set of data, institutional investor's level of transaction cost (0.05%) is assumed for estimation of return. All standard errors are Heteroskedasticity-robust standard errors.

Frequency	States		Coefficient	SE	t-stat	p-value
5 min	3 states	Intercept	-0.00076	0.00155	-0.48735	0.62779
		slope	2.35911	0.65034	3.62749	0.00059
	4 states	Intercept	-0.00081	0.00174	-0.46607	0.64286
		slope	2.18861	0.73122	2.99309	0.00401
	5 states	Intercept	0.00269	0.00213	1.26325	0.21139
		slope	2.56968	0.89342	2.87624	0.00556
10 min	3 states	Intercept	0.00052	0.00226	0.22982	0.81902
		slope	2.74045	0.94916	2.88723	0.00540
	4 states	Intercept	0.00076	0.00199	0.38166	0.70406
		slope	2.66650	0.83448	3.19541	0.00223
	5 states	Intercept	-0.00091	0.00221	-0.41270	0.68130
		slope	2.42797	0.92477	2.62547	0.01096
30 min	3 states	Intercept	0.00068	0.00191	0.35728	0.72214
		slope	2.58314	0.79908	3.23264	0.00199
	4 states	Intercept	-0.00027	0.00188	-0.14424	0.88580
		slope	2.32965	0.78866	2.95394	0.00448
	5 states	Intercept	0.00137	0.00183	0.74937	0.45656
		slope	2.76312	0.76797	3.59796	0.00065

Source: Author's calculation

Table 12: Jenson's Alpha (Continuous HMM, KOSPI50, 0.05% transaction cost)
The following tables compare the trading performance of the continuous Hidden Markov Model in KOSPI50 over the horizon of 1st November 2016 to 31st January 2017. The prediction method of mean return (please see method I of section 4.4.4.3) is used for generating signal. The independent variable is the excess daily return of portfolio and the dependent variable is the excess daily market return of KOSPI. The adjusted daily 3-month Korea Interbank Offered Rate is used as a proxy to risk-free rate. For this set of data, institutional investor's level of transaction cost (0.05%) is assumed for estimation of return. All standard errors are Heteroskedasticity-robust standard errors.

Frequency	States		Coefficient	SE	t-stat	p-value
5 min	3 states	Intercept	0.00299	0.00205	1.45514	0.15076
		slope	0.46458	0.22021	2.10973	0.03899
	4 states	Intercept	0.00038	0.00191	0.19750	0.84409
		slope	0.27108	0.28493	0.95138	0.34517
	5 states	Intercept	0.00048	0.00147	0.32415	0.74693
		slope	0.63291	0.27200	2.32686	0.02331
10 min	3 states	Intercept	-0.00022	0.00197	-0.11374	0.90982
		slope	0.34025	0.30512	1.11514	0.26916
	4 states	Intercept	0.00007	0.00197	0.03491	0.97227
		slope	0.19578	0.22615	0.86571	0.39004
	5 states	Intercept	0.00105	0.00230	0.45817	0.64846
		slope	0.37937	0.26423	1.43577	0.15618
30 min	3 states	Intercept	0.00116	0.00208	0.55657	0.57986
		slope	0.01775	0.21610	0.08216	0.93479
	4 states	Intercept	0.00065	0.00237	0.27600	0.78348
		slope	-0.05843	0.18161	-0.32174	0.74875
	5 states	Intercept	0.00168	0.00245	0.68566	0.49552
		slope	-0.11639	0.19171	-0.60712	0.54602

Source: Author's calculation

CONCLUSION AND RECOMMENDATION

Performance of the discrete and continuous models

In this study, we propose one approach for the discrete models (predict by probability) and two approaches (predict by mean return or probability) for the continuous models to generate the signals to predict the market movement.

For all frequencies, all models with different number of states, the results from the continuous models show lack of predictability; as a result, the trading strategy is not able to generate any abnormal return. On the other hand, the discrete models are able to achieve varied degrees of hit ratio on different frequency and market liquidity and as a result the profitability of the strategy is highly dependent on predictability of the models.

A plausible explanation of the failure of the continuous models to generate meaningful results in both approaches is due to the assumption of distribution. As discussed in section 4, due to the non-normality of intra-day data, this study incorporates the 3 components Gaussian Mixture Model as the distribution function to describe the observable emission of the Hidden Markov Models. However, the BIC score of the GMM models are better, but not only marginally better. In light of this concern, the estimated models are unable to obtain a suitable distribution that best describes the emission. In contrast, the discrete models do not require explicit assumption but requires an appropriate method for discretizing the data. In this study, we have presented a trivial method for discretizing the stock price movement and order imbalance ratio, but there is still lot of rooms for improvement which might benefit the model.

Due to the fact that the continuous models fail to generate meaningful result, in the following section, we focus the discussion of the study on the result of discrete models.

Implication on the performance of the models in different frequencies

As discussed in section 5, the predictability of the models decreases as we lower the frequency (from 5 minutes to 30 minutes). The decrease in predictability takes in two forms in this study; first as the frequency decreases, the observed average hit ratio

decreases. Second, as the frequency decreases, the models are less consistent and confident in generating signals. For instance, moving from 5 minute to 10 minute, the total number of signals generated from the models decrease at a factor of 5 instead of the expected number of 2.

The decrease in confidence of the models in making a prediction seems to be consistent with the previous literature by Chordia et al (2005). The predictability of stock price tends to disappear over a short horizon due the price are already adjusted to the new information. Therefore, as time passes, the stock price movement tends to random walk; and as a consequence, the models are unable to make prediction because no clear patterns are observed from the data.

Expectedly, due to better predictability in high frequency data, the trading strategy gain the highest abnormal return and Sharpe's Ratio for the highest frequency case in the Thai market. Due to the lower predictability of the models in the Korean market, the strategy is only able to generate abnormal return when there is no transaction presents. We further discuss the possible explanation on why the models achieve lower predictability in the Korean market in the following section.

Implication on the performance of the models in different markets

Detailed from the study by Chordia et al (2008), the market liquidity enhances the speed of convergence to market efficiency. As a result, in a more liquid market, the predictability of stocks tends to disappear. Therefore, to assess the applicability of the models under different liquidity environment, we test our strategy in both Korean and Thai stock market, where the market liquidity in the Korean market is relatively higher.

The result seems to be consistent with the previous literature, the predictability of the models are relatively better in terms of both hit ratio and number of signals generated in Thai stock market in comparison to Korean stock market given that the daily trading time in Korean stock market is longer than Thailand; and due to the lower predictability, even at the highest frequency, the strategy is not able to generate abnormal return at institutional level of transaction cost and achieves abnormal return only at case when transaction cost does not exist.

One logical explanation on why market efficiency is enhanced in a more liquid regime is the narrower bid-ask spread. With low bid-ask spread, the barrier for investors to take advantage of momentarily mispricing of financial assets; and as a result, the speed adjustment of information into price is enhanced.

Recommendation for further study

Based on the observations from this study, the discrete Hidden Markov Model is the more appropriate model than the continuous version due to the benefit of no requirement on assumption of distribution. However, there are rooms for improving the methods of discretizing data; in this study, we propose a simple method for discretizing with order imbalance ratio and stock price movement. Due the process of forming the order imbalance ratio, the information of absolute size of order imbalance is lost; the relevance of the absolute size of order imbalance remains untouched. Therefore, it is recommended for future study to formulate a method to incorporate the absolute size of order imbalance into the discretization process.

In addition to the improvement on the models, future study can also consider increasing the frequency of data. For this study, we set the highest frequency of data to be 5 minutes in order to avoid occurrences of no trade. However, for purpose of making gain from trading, the pursuer of opportunity should aim for higher frequency data.

REFERENCES

Chordia, T., Roll, R., & Subrahmanyam, A. (2002). Order Imbalance, liquidity and market returns. *Journal of Financial Economics*, 3-28.

Chordia, T., Roll, R., & Subrahmanyam, A. (2005). Evidence on the speed of convergence to market efficiency. *Journal of Financial Economics*, 271-292.

Chordia, T., Roll, R., & Subrahmanyam, A. (2008). Liquidity and market efficiency. *Journal of Financial Economics*, 249-268.

Fama, F. (1970). Efficient Capital Market: A Review of Theory and Empirical Work. *Journal of Finance*, 383-417.

Hassan, R. (2005). Stock market forecasting using hidden Markov model: a new approach. *5th International Conference on Intelligent Systems Design and Applications*, 192-196.

Hassan, R. (2009). A combination of hidden Markov model and fuzzy model for stock marke forecasting. *Neurocomputing*, 3439–3446.

Hassan, R. (2013). A HMM-basedadaptivefuzzyinferencesystemforstockmarketforecasting. *Neurocomputing*, 10-25.

Patrik, I., & Conny, J. (2008). Algorithmic Trading: Hidden Markov Models on Foreign Exchange Data. *Master Thesis, Department of Mathematics, Linkopings University*.

Satish, R., & Jerry, H. (2010). *Analysis of Hidden Markov Models and Support Vector Machines in Financial Applications.* Berkeley: University of California .

Shen, D. (2015, May 27). Order Imbalance Based Strategy in High Frequency Trading. *Master Thesis of Linacre College*.

Tenyakov, A. (2014). Estimation of Hidden Markov Models and Their Applications in Finance. *Electronic Thesis and Dissertation Repository*, 2348.