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

Po-Lin Wu   Wasin Siwasarit

THAMMASAT UNIVERSITY

Motivation

“How long does it take to remove serial dependence over a daily horizon? The pattern of intra-day serial dependence reveals that it takes more than 5 minutes but less than 60 minutes”

Tarun Chordia (2005)

Evidence on the speed of convergence to market efficiency

Journal of Financial Economics

Motivation

“Prices were closer to random walk benchmarks in the more liquid regime than in others one … these findings indicate that liquidity stimulates arbitrage activity, which, in turn, enhances market efficiency”

Tarun Chordia (2008)

Liquidity and market efficiency

Journal of Financial Economics

Outline

- Objectives of this study
- Scope of data of this study
- Literature Review
- Methodology
- Contribution
- Limitation
- Result
- Conclusion and Recommendation

Objectives of this Study

(1) Attempt to create a model to capture the states of order imbalance selected stocks from SET50 and KOSPI50 in a consistent and confident manner

(2) Build and simulate trading strategies by using the signals generated from proposed models

(3) Compare the result between markets and with traditional buy-and-hold strategies

Scope of Data in this Study

- Data: closing price, bid size, ask size
- Data is limited to stocks in the SET50 Index and KOSPI50 Index
- Scope of data: 1^{st} Nov 2016 to 31^{st} Jan 2017
- Frequency of data: Intra-day data
  - Interval: 5 minutes, 10 minutes and 30 minutes

Data is limited to stocks in the SET50 Index and KOSPI50 Index

- Stocks Selection Criteria
  - Stocks that are consistently listed on SET50 and KOSPI50 Index during the period from 1st January 2012 to 31st July 2016
  - Select 10 stocks with highest 250 days average daily volume turnover
- Data for training and initial testing
  - Initial Training: 1st Aug 2016 to 31st Sep 2016
  - Back testing: 1st Nov 2016 to 31st Jan 2017

Scope of data:
15^{th} Aug 2016 to 31^{st} Jan 2017

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

Scope of data:
15^{th} Aug 2016 to 31^{st} Jan 2017

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

Literature Review

Author	Model	Input Data	Output	Result/Conclusion
Thanpun, P (2012)	Support Vector Machine	Total Return Index, Technical and macroeconomic indicators (Weekly)	Next weekly closing price of total return index	Overall Predicting Power: Correct sign detection: 60~75%
	Probabilistic Neural Network	Total Return Index, Technical and macroeconomic indicators (Weekly)	Next weekly closing price of total return index	Profitability (yearly average): US Case: ~69% at best Thai Case: ~46% at best Hong Kong Case: ~47% at best
	Back Propagation Neural Network	Total Return Index, Technical and macroeconomic indicators (Weekly)	Next weekly closing price of total return index	
Sittipong, S (2013)	Neural-Fuzzy System	SET50 Index and technical indicators (Daily)	Next day closing price	Predicting Power: 51.84% on predicting correct signal

Literature Review

Author	Model	Input Data	Output	Result/Conclusion
Hassan R, (2005)	Continuous HMM (Gaussian Mixture Model)	Open, Close, Intra-day High, Intra-day low price (Daily)	Next day closing price	MAPE/RMSE score R-square
Patrik, I., & Conny, J. (2008)	Discrete HMM with k-mean initialization	EURUSD exchange rate Economic factors (Daily)	Next day exchange rate movement	Model is able to generate profit, but is lack of stability. Too many factors will cause instability to the model.
	Continuous HMM with k-mean initialization	EURUSD exchange rate Economic factors (Daily)	Next day exchange rate movement	
Hassan R, (2009)	HMM with Fuzzy Logic	Open, Close, Intra-day High, Intra-day low price (Daily)	Next day closing price	MAPE/RMSE score
Hassan R, (2012)	HMM based Adaptive Interference Fuzzy System	consecutive weekly stock index (Weekly)	Next weekly closing price	MAPE/RMSE score

Methodology

Order Imbalance

3 indicators from previous literature (Chordia et al, 2005)

1. Number of buy order less number of sell order
2. Number of buy-initiated shares purchased less number of seller-initiated shares sold
3. Dollars paid by buy initiators less dollars received by sell initiators

Order Imbalance

- In this study, we define the order imbalance indicator as:

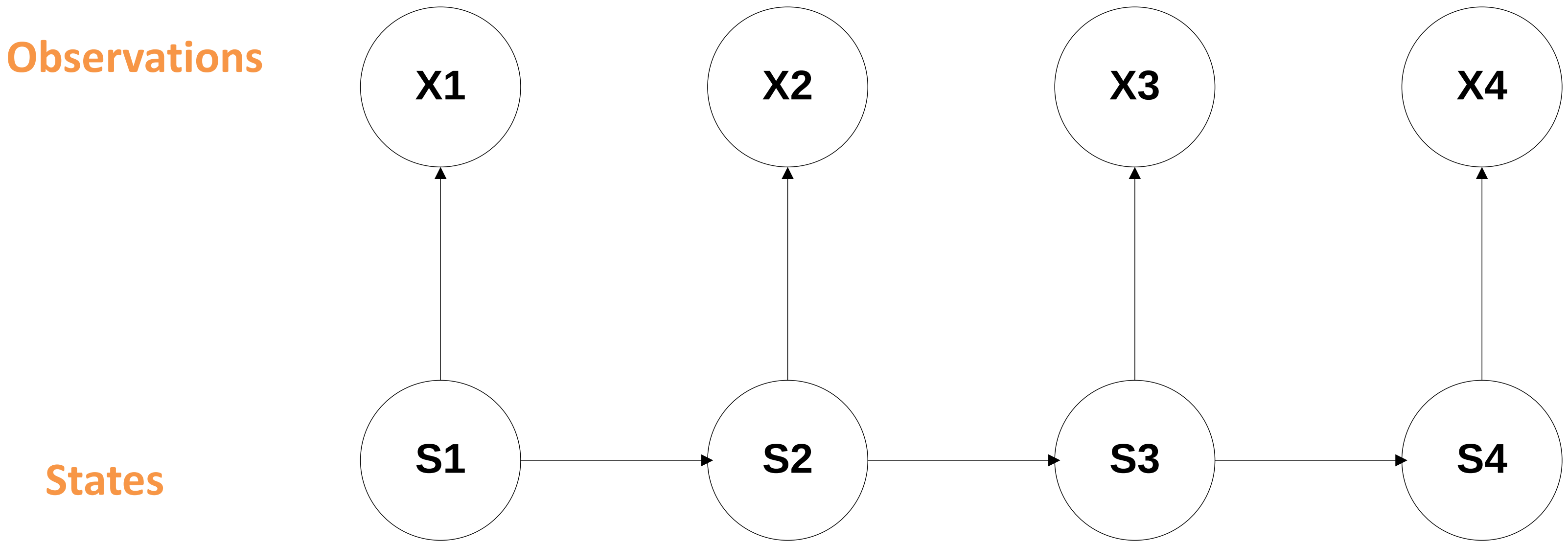
$$\text{OIR} = \frac{V_B}{V_A + V_B}$$

$$V_B = size\ of\ bid\ order\ at\ best\ price$$

$$V_A = size\ of\ ask\ order\ at\ best\ price$$

Hidden Markov Model

Basic Idea:
Hidden Markov Model



Three Fundamental Problems

- **Evaluation Problem: Evaluate how well the model is able to explain the observations**
- **Decoding problem: find the most probable current state**
  - **Viterbi algorithm**
- **Learning problem: find the model that best explain the observations**
  - **Number of states**
  - **Rolling Window: 15 trading days for 5 minutes, 30 trading days for 10 and 30 minutes**

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

Number of states

- This study considers models from 3 states to 5 states
- Three fundamental states:
  - Asset is not adjusted to positive information
  - Asset is not adjusted to negative information
  - Asset price is in equilibrium
  - Other possible states

Generating trading signal

- We plan to generate trading signal by 2 approaches:
  - **Discrete Case**
  - **Continuous Case**

Data Discretization

- **Before the model training, data needs to be discretized:**

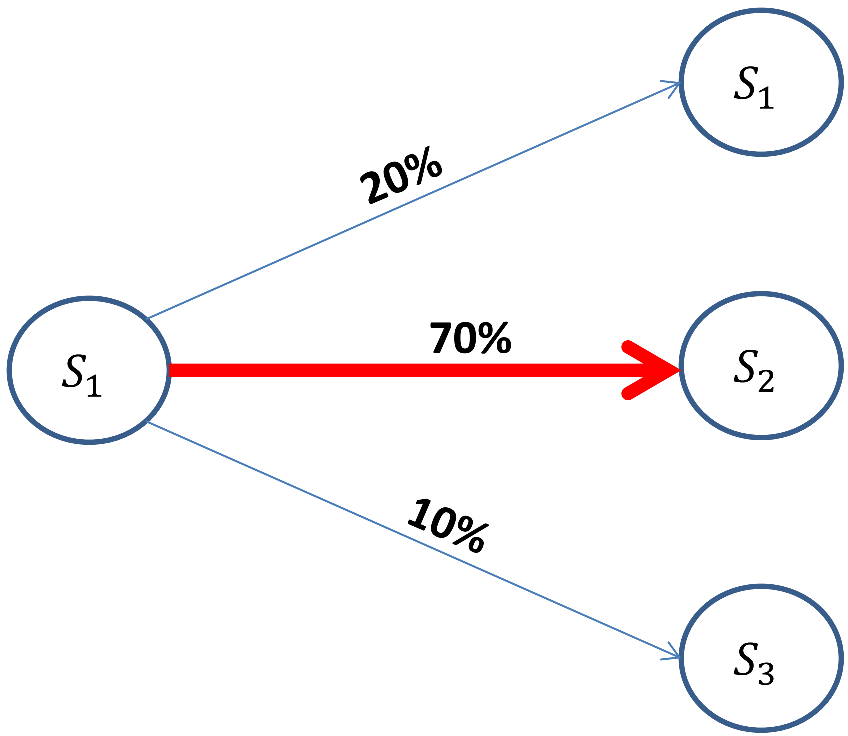
Symbol	Return	Order Imbalance Ratio
1	$0\% <$	$< 25\%$ quantile
2	$0\% <$	25% quantile $\leq$ OIR $\leq 0.75\%$ quantile
3	$0\% <$	OIR $> 75\%$ quantile
4	$\geq 0\%$	$< 25\%$ quantile
5	$\geq 0\%$	25% quantile $\leq$ OIR $\leq 0.75\%$ quantile
6	$\geq 0\%$	OIR $> 75\%$ quantile

Noted: the percentile is computed by averaging the percentile of each selected stocks during the initial training period

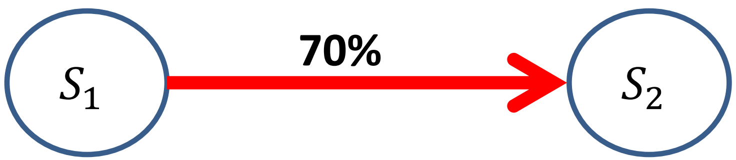
Data Discretization

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

Generating trading signal



Case1 :Discrete Trading signal



Probability of emission:

Symbol	Return	Probability
1	$0\% <$	5%
2	$0\% <$	5%
3	$0\% <$	5%
4	$\geq 0\%$	30%
5	$\geq 0\%$	30%
6	$\geq 0\%$	25%

$$P(q_{t+1} = S_2 | q_t = S_1) P(r > 0 | S_2)$$

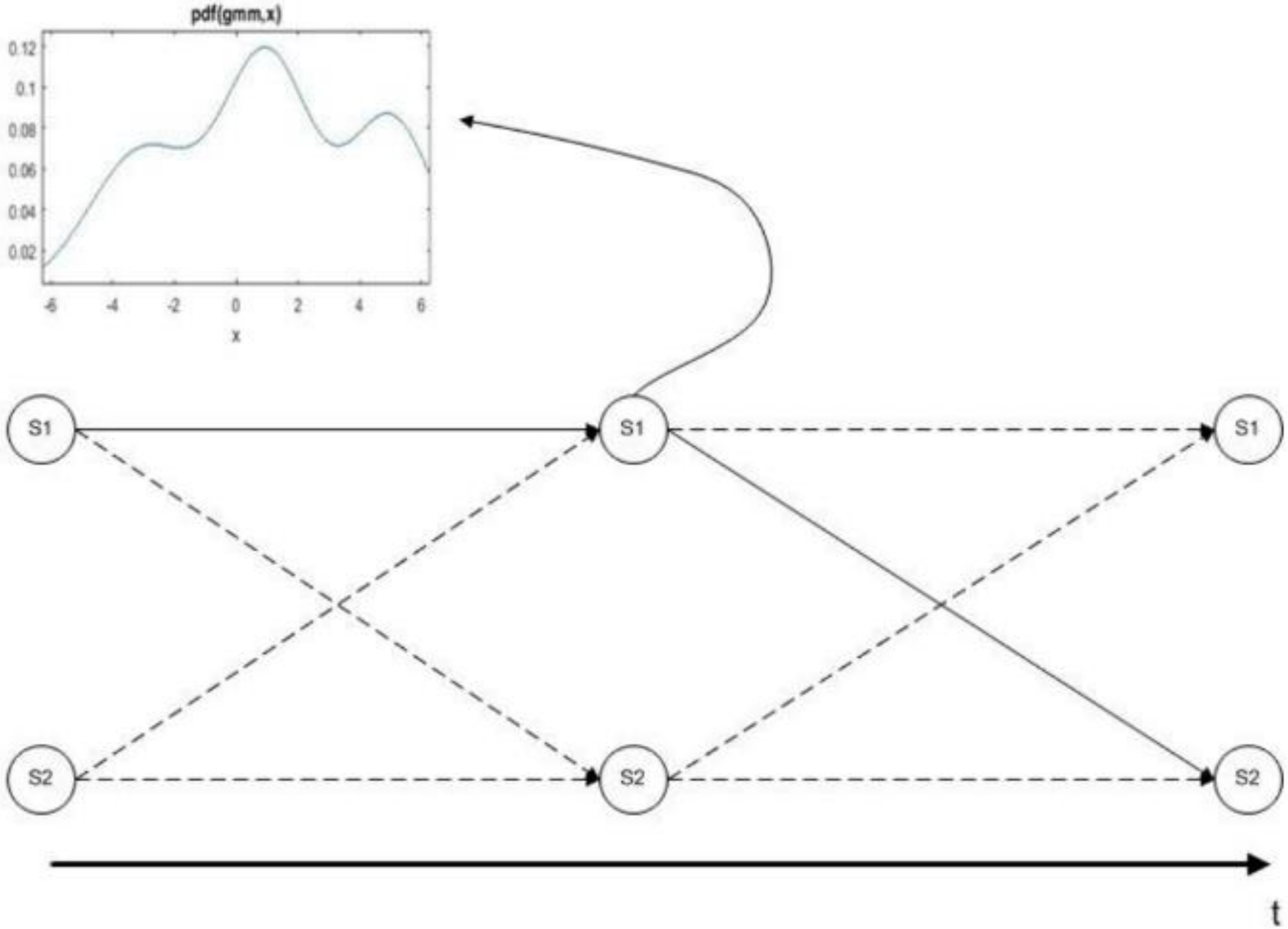
If $P \geq threshold$, then signal is generated

Intuition behind threshold

- This study aims to capture the price movement in a confident manner:
  - **Confident in transition: 90%**
  - **Confident in observing the desire movement: 90%**

- The joint-probability gives us an approximate number of 80% for our threshold value

Trading signal (Continuous)



Case2: Continuous Trading Signal

- Intra-day data is not normally distributed

Table 1: Summary statistics of SET50: return of 5 Minute data

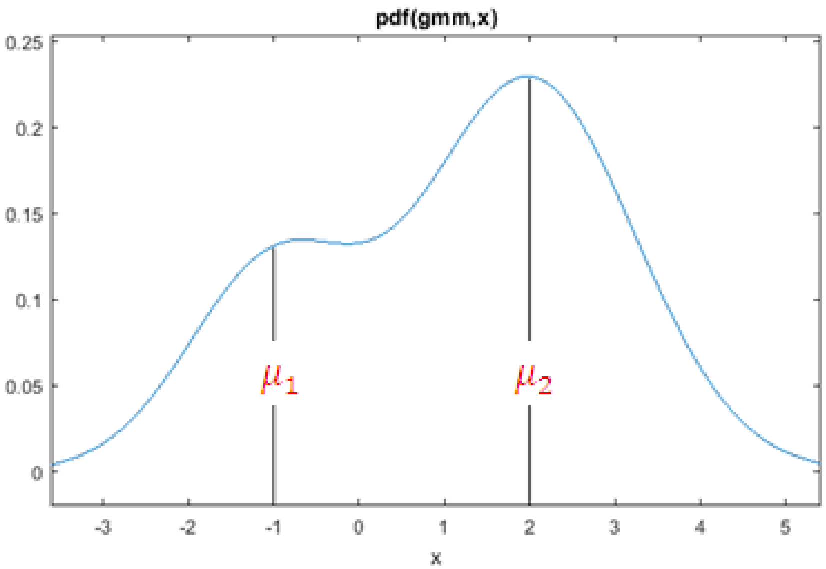
Ticker	μ	σ	Skewness	kurtosis
ADVANC.BK	0.00000	0.00287	-0.57476	35.13078
BANPU.BK	0.00004	0.00448	-0.19062	13.17991
BCP.BK	0.00003	0.00504	-0.00774	4.37384
CPF.BK	-0.00003	0.00607	-0.19439	6.49320
DTAC.BK	0.00006	0.00577	-0.43297	14.74661
IRPC.BK	0.00001	0.00420	-0.58875	15.65829
IVL.BK	0.00005	0.00575	-0.40375	10.40947
PTTEP.BK	0.00004	0.00293	0.20801	22.99403
TCAP.BK	0.00003	0.00422	0.15091	4.66762
TRUE.BK	-0.00002	0.00538	-0.01786	13.59031

Trading signal (Continuous)

- Gaussian Mixture Model is more capable to describe the intra-day data

Ticker	Number of Components	Negative Loglikelihood	AIC	BIC
034220.KS	1	-29937.7	-59865.4	-59831.6
	2	-30162.5	-60302.9	-60228.4
	3	-30214.2	-60394.4	-60279.3
	4	-30215.9	-60385.9	-60230.1
	5	-30245.7	-60433.3	-60237
066570.KS	1	-31341.4	-62672.8	-62639
	2	-31374.8	-62727.7	-62653.2
	3	-31382.1	-62730.3	-62615.2
	4	-31438.5	-62831.1	-62675.3
	5	-31472.6	-62887.2	-62690.8
051910.KS	1	-30305.8	-60601.6	-60567.7
	2	-30533.6	-61045.2	-60970.7
	3	-30577.5	-61121.1	-61006
	4	-30575.8	-61105.7	-60949.9
	5	-30579.6	-61101.2	-60904.8

Trading signal (Continuous)

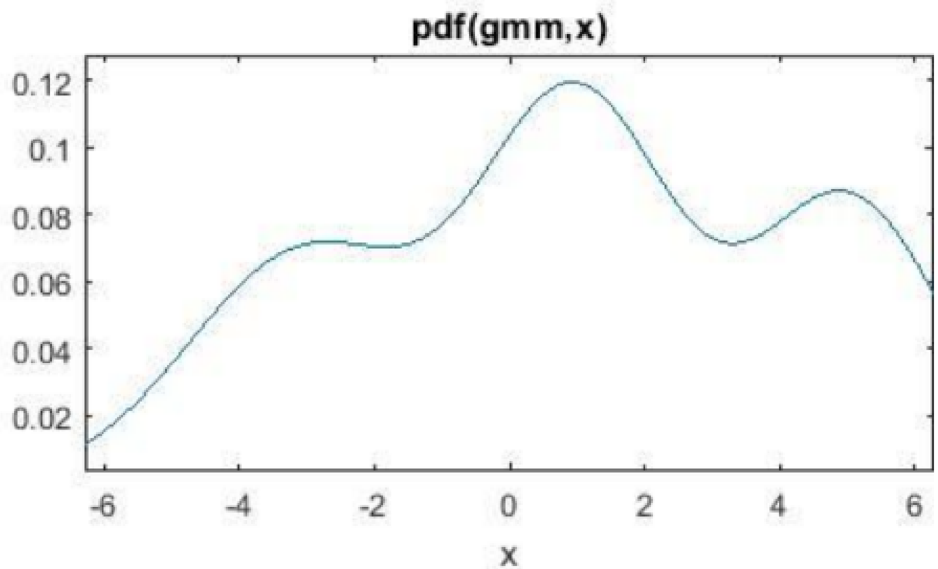


Approach I: By using first moment

$$E(r) = \sum_{m=1}^{M} w_i \mu_i$$

trading signal is then generated based on calculated expected return

Trading signal (Continuous)



Approach II: By using probability

$$\sum_{i=1}^{M} w_i P(lagged\ return \geq d | G(\mu_i, \Sigma_i))$$

trading signal is then generated based on calculated probability

Trading strategy

1. **If current period is the end of the day, then re-train the model for each stocks**

2. **Generate a list of stocks that we should enter long position**

3. **Liquidate any stocks that are not in the list**

4. **If there is any remaining wealth, allocate wealth equally to all stocks in the list**

5. **Proceed to next period**

The study will consider both with/without transaction cost cases (0.05% and 0.1% bi-directional)

Performance evaluation

- Hit ratio: how well the signal is able to predict positive return for each individual stock

$$Hit\ Ratio = \frac{h}{n}, where \begin{cases} h = number\ of\ signal\ that\ predict \\ \quad positive\ return\ correctly \\ n = number\ of\ signal\ generated \end{cases}$$

- t-test: $H_0: Hit\ ratio = 0.5, H_a: Hit\ ratio > 0.5$

Performance evaluation

- **Benchmark: SET and KOSPI total return Index**
- **Sharpe ratio**
- **Jenson’s Alpha**

Result

Predictability

Hit ratio: SET50, Discrete case

	3 states	4 states	5 states
5 min	78.61%	80.60%	83.38%
10 min	72.21%	72.41%	67.15%
30 min	71.81%	62.89%	62.20%

Observations

- As frequency decreases, the predictability of the signals decreases.

- At highest frequency, model with higher number of states achieve higher hit ratio in comparison.

Hit ratio: KOSPI50, Discrete case

	3 states	4 states	5 states
5 min	72.71%	58.93%	71.57%
10 min	66.16%	69.61%	59.41%
30 min	60.00%	43.04%	45.04%

Observations

- Similar pattern is obersved, as frequency decreases, the hit ratio decreases

- Compare to Thai market:
  - hit ratio is lower across all frequency and models
  - Though having longer trading hours, number of signals is significantly lower

Hit ratio: SET50, Continuous case

	3 states	4 states	5 states
5 min	50.06%	49.91%	49.88%
10 min	49.53%	49.33%	49.48%
30 min	49.62%	50.09%	49.65%

Hit ratio: KOSPI50, Continuous case

	3 states	4 states	5 states
5 min	49.96%	50.10%	50.04%
10 min	49.85%	49.61%	49.61%
30 min	49.25%	49.19%	48.66%

Observation

- The signs show no predictability
  - In all cases, we fail to reject to null hypothesis that the hit ratio is no better than random guessing
- Possible cause of meaningless result: assumption of probability distribution

Observation

- Extreme Kurtosis property of intra-day data:

Ticker	μ	σ	Skewness	kurtosis
034220.KS	0.00001	0.00262	-0.19991	24.20587
066570.KS	0.00003	0.00212	0.61680	15.73020
051910.KS	0.00002	0.00237	-1.27951	42.47450
005490.KS	0.00003	0.00254	0.74383	23.81689
006400.KS	0.00003	0.00282	-0.35090	25.74632
009150.KS	0.00002	0.00209	2.92288	76.95801
010140.KS	0.00002	0.00289	1.00025	17.88761
000880.KS	0.00000	0.00215	0.88572	17.47675
000720.KS	0.00002	0.00243	-0.29044	20.36857
009540.KS	0.00000	0.00354	3.27655	96.11003

Profitability

Sharpe's Ratio
SET50, Discrete case

	3 states	4 states	5 states	Benchmark
5 min	2.36	4.83	5.51	1.62
10 min	4.83	0.37	0.48	1.62
30 min	2.86	0.21	0.47	1.62

Be noted: 0.05% bi-directional transaction cost assumed

Jenson's Alpha
SET50, Discrete case

	3 states	4 states	5 states
5 min	2.87%***	3.57%***	4.73%***
10 min	1.06%***	0.14%	0.41%
30 min	0.77%***	0.05%	0.06%

Noted: 0.05% bi-directional transaction cost assumed. Number of stars represent the level of significance, with *** represents 99% significance, ** represents 95% significance. 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.. All standard errors are Heteroskedasticity-robust standard errors.

Sharpe's Ratio
KOSPI50, Discrete case

	3 states	4 states	5 states	Benchmark
5 min	-0.51	1.04	0.10	0.62
10 min	-0.78	-1.42	-1.26	0.62
30 min	-0.66	-1.76	-1.14	0.62

Jenson's Alpha KOSPI50, Discrete case

	3 states	4 states	5 states
5 min	-0.09%	0.05%	0.03%
10 min	-0.10%	-0.28%***	-0.17%**
30 min	-0.04%	-0.08%	-0.17%*

Noted: 0.05% bi-directional transaction cost assumed. Number of stars represent the level of significance, with *** represents 99% significance, ** represents 95% significance. 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. All standard errors are Heteroskedasticity-robust standard errors.

Sharpe's Ratio SET50, Continuous case

	3 states	4 states	5 states	Benchmark
5 min	0.01	-0.07	4.45	1.62
10 min	0.50	0.63	-0.06	1.62
30 min	0.80	0.52	1.30	1.62

Jenson's Alpha
SET50, Continuous case

	3 states	4 states	5 states
5 min	-0.08%	-0.08%	0.05%
10 min	0.05%	0.08%	-0.09%
30 min	0.07%	-0.03%	0.14%

Noted: 0.05% bi-directional transaction cost assumed. Number of stars represent the level of significance, with *** represents 99% significance, ** represents 95% significance. 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.. All standard errors are Heteroskedasticity-robust standard errors.

Sharpe's Ratio
KOSPI50, Continuous case

	3 states	4 states	5 states	Benchmark
5 min	1.08	0.17	0.84	0.62
10 min	-0.02	0.06	0.40	0.62
30 min	0.23	0.11	0.30	0.62

Noted: 0.05% bi-directional transaction cost assumed

Jenson's Alpha KOSPI50, Continuous case

	3 states	4 states	5 states
5 min	0.30%	0.04%	0.05%
10 min	-0.02%	0.01%	0.11%
30 min	0.12%	0.07%	0.17%

Noted: 0.05% bi-directional transaction cost assumed. Number of stars represent the level of significance, with *** represents 99% significance, ** represents 95% significance. 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. All standard errors are Heteroskedasticity-robust standard errors.

Conclusion

Performance of models

- With inappropriate assumption of probability distribution of observations, the continuous model failed.

- Since discrete model require no assumption on distribution of observations, the model is able to produce a meaningful result

Effect of frequency on performance of models

- Observed from both market, the higher the frequency, the higher the predictability. The same goes for profitability.

- Consistent with the literature by Chordia et al (2005): Price adjustment to information occurs on the intra-day level and predictability tends to dispear when frequency decreases.

Effect of market liquidity on performance of models

- In market with higher liquidity, the model is less consistent and confident:
  - Generate less signals
  - Achieve lower hit ratio

- Consistent with previous literature (Chordia et al, 2008): In a high liquidity environment, the cost of trading is lower (ex. bid-ask spread). Hence, investors have more incentives to exploit the deviation of asset price from equilibrium. This enhances the speed of price adjustment to new information.

Contribution

- Formulating trading strategies for institutional traders
- Provide insight to efficiency of stock market in different market liquidity and intra-day frequency

Limitation

- No short selling
- Only consider stocks that are satisfied with discussed criteria
- The back-testing period is limited to 3 months
- Trade will be executed base on mid-point closing price but not actual the bid-ask price

Recommendation

- Focus on data of higher frequency
- Re-consider the data discretization method
- Re-consider the method of constructing order imbalance indicator

Source : https://www.maanavan.com/wp-content/uploads/2017/04/question-and-answer-images-clipart-panda-free-clipart-images-EodNcb-clipart.png