On the Evaluation of Intraday Market Quality in the Limit-Order Book Markets : A Collaborative Filtering Approach

Hayashi, Takaki / Takahashi, Makoto

(出版者 / Publisher)
法政大学イノベーション・マネジメント研究センター
(雑誌名 / Journal or Publication Title)
法政大学イノベーション・マネジメント研究センター ワーキングペーパーシリーズ
(巻 / Volume)
239
(開始ページ / Start Page)
1
(終了ページ / End Page)
32
(発行年 / Year)
2021-06-08

WORKING PAPER SERIES

Takaki Hayashi and Makoto Takahashi

On the Evaluation of Intraday Market Quality in the Limit-Order Book Markets: A Collaborative Filtering Approach

June 8, 2021

No. 239

The Research Institute for Innovation Management, HOSEI UNIVERSITY

On the evaluation of intraday market quality in the limit-order book markets: A collaborative filtering approach*

Takaki HAYASHI† ‡ § and Makoto TAKAHASHI¶

Abstract

This study proposes a methodology for evaluating market quality of individual stocks in the high-frequency domain (short term) by applying a recommender system that has become ubiquitous in our daily lives, especially when running internet apps.

In the first place, it is not easy to evaluate market quality such as the "true" liquidity of individual stocks. In particular, in situations where liquidity for a short-time period is to be evaluated using high-frequency data, the lack of observations can become severe for numerous stocks. Since stocks that have exhibited similar behavior in the past are expected to perform so in the future as well, one can expect that *collaborative filtering*, which is nowadays the main approach of *recommender systems*, can work effectively for the market quality measure "estimation" problem for stocks. However, in some occasions, "standard-type" collaborative filtering methods may not work well, especially when data sparsity (or scarcity) is severe.

Specifically, in this paper we adopt a regression-based latent factor model (*RLFM*), a "hybrid-type" collaborative filtering proposed by Agarwal and Chen (2009). It has a hierarchical linear structure designed to address the so-called "cold-start problem" in the recommender systems literature.

In this study we take on liquidity and volatility as market quality measures. The liquidity measure used in this paper is the *ILOBS* (Inverse Limit Order Book Slope) proposed by Deuskar and Johnson (2011). For volatility, we adopt the realized volatility based on tick-by-tick mid-quote prices.

In order to investigate the effectiveness of the method in consideration, empirical analysis was performed using high-frequency limit-order book data from the Tokyo Stock Exchange. The data period is the first three months of the year 2019, with which regularly-spaced five-minute aggregate datasets were formed. The explanatory variables in the regression term are six variables related to observed market activities of individual stocks such as logarithmic return and share volume, three variables related to the static attributes of individual stocks such as whether it is an ingredient of the Nikkei 225 Index, and the industry category it belongs to. There are also time polynomial terms up to the order of 6 to capture the average movements of the whole market along the time axis. As a result of the empirical analysis, various characteristics that characterize market quality were identified from the estimated regression coefficients obtained by fitting the RLFM model to the training dataset. The results suggest that our approach is robust to the degree of sparsity (or scarcity) and flexible enough to deal with various data environments. There was room for improvement of the methodology in terms of prediction accuracy.

Keywords: Collaborative filtering, high-frequency data, latent factor models, limit-order book, liquidity, market quality, matrix completion, volatility, recommender systems.

1 Introduction

With the rapid development of information and communication technology (ICT), the "third-generation" AI boom has arrived. Digital transformation (DX) has spread in corporate organizations, and cross-tech (x-Tech) companies that provide customers with new products and services by applying ICT technologies have

*This is a post-peer-review, pre-copyedit version of an article published in *Japanese Journal of Statistics and Data Science*. The final authenticated version is available online at: http://dx.doi.org/10.1007/s42081-021-00116-0.

†Graduate School of Business Administration, Keio University. hayashi.kbs@gmail.com

‡Finance Program, Graduate School of Management, Tokyo Metropolitan University.

§CREST, Japan Science and Technology Agency.

¶Faculty of Business Administration, Hosei University. m-takahashi@hosei.ac.jp

emerged. Many FinTech companies have also appeared in the financial industry, providing more convenient services, offered at lower costs, all of which has made the industrial landscape change drastically. Under such circumstances, research on the application of AI/machine learning technology in various applications of finance is being pursued by practitioners and academic researchers.

This study examines a methodology for evaluating market quality of individual stocks traded on the limit-order book markets by using a recommender system approach. *Recommender systems* are "software tools and techniques that provide suggestions for items that are most likely of interest to a particular user.... The suggestions relate to various decision-making processes, such as what items to buy, what music to listen to, or what online news to read." Mainly targeted are those individuals who "lack the sufficient personal experience or competence in order to evaluate the potentially overwhelming number of alternative items that a website, for example, may offer" (Ricci et al. (2015)). Today we are indeed surrounded by various recommender systems that can make our daily lives more convenient and comfortable; e.g., Amazon.com, YouTube, Netflix, Spotify, LinkedIn, Facebook, and so forth.

Beginning in the mid 90's, academic research on recommender systems has progressed greatly in the current century as practical applications have been wide spreading.[1]

One primary goal of recommender systems is to predict the unobserved (or missing) value of a target user's rating on a particular item. It is formulated as a *matrix completion problem* for a user-item combination matrix, the (i, j)-th entry of which represents the user i's rating on the item j; the entry is missing if she has yet experience nor provided feedback on the item.

Recommender systems have two major approaches: *content-based* and *collaborative filtering*. The latter is based on evaluation and usage history without using exogenous information about items or users. It is a method of recommending items to individual users based on the patterns contained in the past behaviors of the users in the dataset.

Collaborative filtering methods and its variants are nowadays widely implemented for recommender systems. An epoch-making event that stimulated and advanced collaborative filtering methods was *The Netflix Movie Challenge* launched in 2006 and ended in 2009. In this contest, more than 40,000 teams of data scientists from 186 countries over the world participated with a newly developed algorithm to compete for prediction accuracy in order to improve the company's system called *CineMatch* algorithm for recommending movies to their customers.[2]

There are two types of collaborative filtering commonly used in practice, *memory-based methods* and *model-based methods*. In memory-based methods, ratings of target users are predicted on the basis of their neighborhoods. The neighborhoods are defined in terms of either user-user distance or item-item distance. *User-based methods* firstly identify peer users who are the closest to a target user regarding ratings of the items they have already made and then predict ratings for the unobserved items of the target user by computing weighted averages of the ratings of the peer group. Similarity functions are defined between the users, i.e., between the rows of the user-item rating matrix to find similar users. In the meantime, *item-based methods* firstly identify a set of items that are most similar to a target unobserved item. Then, the ratings in the item set, given by the target user, are used to predict the rating of the target item. Similarity functions are defined between the items, or the columns of the user-item rating matrix to find similar items.

In model-based methods, a model to summarize user-item data is constructed *a priori*. In contrast to memory-based, model-based methods do not need to hold the whole dataset on memory. In practice,

[1] According to ACM Digital Library (http://dl.acm.org/), the number of articles whose title contains "recommender systems" has been on the surge; 26 in 1999–2003, 221 in 2004–2008, 943 in 2008–2013, and 1381 in 2013–2019 (cf. Ricci et al. (2015)).

[2] https://web.archive.org/web/20131030190759/http://www.netflixprize.com//index

when dealing with large-scale datasets, that aspect is essential. For dimension reduction of matrices, matrix factorization such as singular value decomposition (SVD) plays a key role. Since user-item rating matrix data is extremely sparse in general, i.e., there are an enormous number of missing values, in the context of recommender systems, dimension reduction and missing value imputation have to be done at once. It turns out that matrix factorization methods can be employed to exploit row and column correlations together to estimate/complete the whole data matrix. *Latent factor models* that rely on a low-rank matrix decomposition are widely used. One supposes that the user i's rating on the item j is expressed as

$$y_{ij} = u_i' v_j + \epsilon_{ij},$$

where $u_i = (u_{il})$ and $v_j = (v_{jl})$ are both $(r \times 1)$-vectors of *latent factors* and ϵ_{ij} is an "error" term. In a context of movie reviews, there are r "genres" of movies, u_{il} indicates "affinity" between user i and genre l, v_{jl} does so between item j and genre l.

See Aggarwal (2016) for further exposition on recommender systems and collaborative filtering.

By seeing analogy of the filling (predicting) of missing consumer movie reviews to that of unobserved market quality measures of individual stocks, a collaborative filtering method can possibly be applied to the behavior of market quality measures for the target stocks to be evaluated. It is expected that the market quality measures of the target stock can be evaluated (predicted) by finding the "neighbors" and suitably aggregating their observable information. So far as we are aware, the application of collaborative filtering to the problem of market quality evaluation is new and hence is worth exploring its potentials.

2 The methodology

2.1 The "estimation" problems of market quality measures

"Market quality" is a concept that expresses the degree to which a securities market functions "properly," that is, the degree to which fair prices are formed in the market and efficient capital allocation is realized through transactions (cf. Yano (2009)). Key elements that can determine (or measure) market quality include transaction costs (bid-ask spread, etc.), execution speed, volatility, liquidity, order fill rate (the number of traded shares divided by the number of ordered shares).[3]

In this paper, we will explore methodology to evaluate intraday market quality by use of high-frequency data. In particular, we will take on *liquidity* and *volatility*, thinking of their importance both in practice and in academic research. So far as we are aware, methodologies for the evaluation of intraday market quality have yet to be well pursued in the literature, compared with those on daily basis.

Liquidity

Among the elements of market quality mentioned earlier, liquidity plays a central role; it is the degree to which a market participant can trade the quantity she wishes to buy or sell in a short time without changing the market price.

Major liquidity indicators used in research and practice include "depth" (limit order quantity present on an order plate), trading volume (contract volume divided by amount), trading turnover (contract volume

[3] According to Nasdaq, market quality "is difficult to determine a common definition amongst market participants. At its core, it can be defined generally as that which gives end customers a fair deal" (source: https://www.nasdaq.com/articles/assessing-market-quality-2015-12-22).

divided by number of shares issued), Kyle (1985)'s λ (Price Impact Indicator), Amihud (2002)'s ILLIO (Illiquidity Indicator) and the forth.

In this study, we treat the evaluation of liquidity *as a statistical problem*. That is, we regard the observed liquidity measure(s) is subject to *contamination due to measurement error* or *random noise*, hence the "true" liquidity is to be "estimated" from the observed quantities.

In this study, we adopt the *ILOBS* (inverse limit order book slope) proposed by Deuskar and Johnson (2011) as a measure of liquidity. It is defined as

$$ILOBS = \frac{\Sigma_{k=1}^{K}\left(P_k^b - m\right)^2 + \Sigma_{k=1}^{K}\left(P_k^a - m\right)^2}{\Sigma_{k=1}^{K}\left(P_k^b - m\right)\left(-CQ_k^b\right) + \Sigma_{k=1}^{K}\left(P_k^a - m\right)CQ_k^a}. \tag{1}$$

Here, $P_k^{\cdot}$ is the k-th best quoted price from the innermost quoted price (i.e., best bid-offer, or BBO) on either side of the limit-order book, while $CQ_k^{\cdot}$ is the cumulative number of shares up to the k-th quoted price from the BBO on the book. a, b are respectively ask quote, bid quote, while m is the mid quote.

ILOBS is a measure of illiquidity. It is designed to capture the expected effect of executable orders on price. A larger value of ILOBS indicates that the market is less liquid; see Figure 2 of Deuskar and Johnson (2011).

Deuskar and Johnson (2011) showed that the price impact (price change due to order) and ILOBS have highly correlated. In other words, even when perfect information on the shape of the whole book is unknown, ILOBS can predict quite well the price impact, hence it is a reasonable liquidity measure for the purpose of this study.

There have been numerous empirical work on price impact for limit-order book, such as Bouchaud et al. (2009), Eisler et al. (2012), Hautsch and Huang (2012), Cont et al. (2014). However, there seem far less empirical studies on ILOBS irrespective of its potential usefulness. Takahashi (2018) is one of such studies.[4]

Assessing the impact of high-frequency trading (HFT) on market quality has been an important research theme since the 2010 Flash Crush occurred (Litzenberger et al. (2012), Hasbrouck and Saar (2013), Menkveld (2013, 2016), Brogaard et al. (2014), Kirilenko et al. (2017), etc.). Establishing a methodology that can possibly evaluate liquidity properly – especially in the high-frequency domain – should be beneficial for studies on the assessment of HFT's impacts on market quality.

Volatility

Volatility is an unobserved state variable that represents variability of an asset price (or the whole market), playing the key role for financial risk management and asset pricing. There is a thick literature on the estimation of volatility with high-frequency data. Since simply applying the celebrated GARCH model to high-frequency data turned out not working well, partly owing to the existence of stylized facts peculiar to high-frequency data such as intraday periodicity (cf., Dacorogna et al. (2001, Sec.8.2.2)), various modeling efforts have been made to accommodate such phenomena, e.g., Andersen and Bollerslev (1997). In the meantime, *realized volatility* has been studied extensively; Andersen et al. (2001) and Barndorff-Nielsen and Shephard (2002) are among early seminal work in the area.

In this paper, we take a different path from most of the existing literature on volatility estimation taking. For the standard approaches such as ARCH/GARCH models and realized volatilities, past price series are

[4] According to Google Scholar, as of January, 22th, 2020, Deuskar and Johnson (2011) has been cited 33 times.

essentially the only inputs to compute volatility estimates. These approaches are inherently infeasible when prices are not observed. Meanwhile, we regard the estimation problem of volatility as an imputation – or filling missing values – problem or a completion problem for a sparse volatility matrix, with the help of supplementary information collected from itself and from others.

As a target quantity, here we utilize the realized volatility based on tick-by-tick mid-quote prices.[5] Mid-quote prices are used because, they change if and only if either of the best quotes change, possibly carrying a new information arrival or perhaps an indication of a change of the fundamental value of the firm. We note that we do not adjust the realized volatility for the so-called market microstructure noise; for one thing mid-quotes are insensitive to bid-ask bounce, a major element of market microstructure noise, and bias adjustments can be computationally burdensome when computation of a great number of stocks takes place for another.

Recently, substantial research activities on sparse estimation have been conducted especially for covariance matrices across multiple assets. Fan and Kim (2018), Kim et al. (2018), Belomestny et al. (2019), for instance. See Fan et al. (2016) for a review on the estimation of large covariance and precision matrices. They have different purposes and approaches to ours, however are worthy of mention here.

In summary, the purpose of this study is to examine a methodological framework by use of high-frequency data for evaluating market quality in high-frequency domains (short-time period) of individual stocks, which works effectively even for issues with relatively less trading activity.

We adopt ILOBS as the liquidity measure and the realized volatility for mid-quotes as the volatility measure throughout the rest of the paper. However, these choices are arbitrary though we believe are reasonable. They are essentially independent from the estimation framework under investigation, hence other choices for market quality measures and the corresponding estimators are possible.

2.2 The model

In this study, we evaluate the two elements of market quality – liquidity and volatility – by using the high frequency limit-order book data. The degree of scarcity can differ across stocks, especially, in the high frequency domain, where missing values frequently occur depending on the stock and its degree of occurrence changes greatly over time of the day and the market condition.

To fill in such missing values, we consider using auxiliary information such as various (micro) attributes of individual stocks and external macro information surrounding the stock market and itself. In particular, we adopt "hybrid-type" collaborative filtering to deal with the so-called "cold-start problem" where the "standard-type" collaborative filtering may not work well.

Specifically, we adopt the regression-based latent factor model (RLFM) proposed by Agarwal and Chen (2009). It combines a latent factor model for the collaborative filtering with a regression model. The former part incorporates the information on correlation with the market quality measures of other stocks and its own past, and the latter complements the missing measures of some stocks at some time intervals by using explanatory variables such as individual stock attributes and macro information.

Let y_{ij} be an objective variable (liquidity measure) for stock i $(= 1, \ldots, M)$ and time j $(= 1, \ldots, N)$. Then, the RLFM consists of two layers.

[5] The proposed framework accommodates alternative volatility measures such as Parkinson (1980) volatility estimator and others discussed in Section 3.4.

- The first layer (equation for the target variable):

$$y_{ij} = x_{ij}'b + \alpha_i + \beta_j + u_i'v_j + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2). \tag{2}$$

- The second layer (equations for the latent factors):

$$\alpha_i = g_0'w_i + \epsilon_i^{\alpha}, \quad \epsilon_i^{\alpha} \sim N(0, \sigma_\alpha^2), \tag{3}$$

$$\beta_j = d_0'z_j + \epsilon_j^{\beta}, \quad \epsilon_j^{\beta} \sim N(0, \sigma_\beta^2), \tag{4}$$

$$u_i = Gw_i + \epsilon_i^{u}, \quad \epsilon_i^{u} \sim N(0, \sigma_u^2 I_r), \tag{5}$$

$$v_j = Dz_j + \epsilon_j^{v}, \quad \epsilon_j^{v} \sim N(0, \sigma_v^2 I_r). \tag{6}$$

x_{ij} is a $(s \times 1)$-vector (observed market activities of individual stocks), w_i a $(p \times 1)$-vector (static attributes of individual stocks), z_j a $(q \times 1)$-vector (overall market movements attached to the time bin of the day), α_i and β_j are (latent) quantities representing stock and time attributes, respectively, and u_i and v_j are both $(r \times 1)$-vectors of latent factors.

As is evident from the appearance, the RLFM is not identifiable. As per a suggestion made in Agarwal and Chen (2016, p. 160), we put the following constraints on the factor values:[6]

$$\sum_i \alpha_i = \sum_j \beta_j = 0, \qquad \sum_i u_i = \sum_j v_j = \mathbf{0}. \tag{7}$$

There remains another non-identifiability, $u_i'v_j = (-u_i)'(-v_j)$. So, switching of the signs of u_i and v_j results in switching of those of the corresponding regression parameters. However, it will not change the value of the log-likelihood in estimation (Agarwal and Chen (2016, p. 171)).

Further, it is worth noting that we need to select the variables more carefully when we apply the RLFM to economic and financial time-series data.[7] Unlike the user-item data, which is a prime target of the most recommender systems and typically a cross-sectional data collecting the information at the one point or period of time, there is a natural ordering of observations and a possible time trend and/or periodicity in the time-series data. In particular, a presence of a time trend/periodicity among the variables in x_{ij}, w_i, and z_j may cause a serious multicollinearity problem in estimating parameters although it should not be a problem in predicting the target variable y_{ij}. We can avoid the multicollinearity problem by selecting variables in x_{ij}, w_i, and z_j carefully and/or by conducting a suitable data processing in advance. In the empirical analyses below, we remove periodicities (along the time $j = 1, 2, \ldots, N$) from x_{ij} in advance and make z_j capture a common periodicity. The estimation methodology of the RLFM as well as the data processing is illustrated in the next section.

Remarks:

- *Dealing with nonlinearity (1):* Agarwal and Chen (2016) formulate the linear regression functions (2)–(5) in a general manner. That is,

$$\alpha_i \sim N(g(x_i), \sigma_\alpha^2), \qquad \beta_j \sim N(h(z_j), \sigma_\beta^2),$$
$$u_i \sim N(G(w_i), \sigma_u^2 I_r), \quad v_j \sim N(H(z_j), \sigma_r^2 I_r),$$

[6] From practical considerations, these constraints are enforced after sampling by subtracting the sample means. For example, after sampling all factors α_i, we compute $\bar{\alpha} = \frac{1}{M}\sum_i \hat{\alpha}_i$ and set $\hat{\alpha}_i = \hat{\alpha}_i - \bar{\alpha}$, for all i. M is the number of users and $\hat{\alpha}_i$ is the posterior sample mean of α_i. See Agarwal and Chen (2016, p. 160).

[7] We thank an anonymous referee for pointing out this issue.

for suitable functions g, h, G, H. Hence, the linear formulation in this study has room for extensions in this regard.

- *Dealing with nonlinearity (2):* In case the response variable y_{ij} is binary, logistic regression models can be adopted in place of the linear regression model at the first layer.

- *Scalability:* One major advantage of the use of a model-based collaborative filtering approach is its scalability. That is, a single model, such as the RLFM considered here, can apply to small-scale to large-scale recommendation problems. For instance, in the case of The Netflix Movie Challenge, 2006–2009, the dataset provided for participants had N=17,770 movies (columns) and M=480,189 customers (rows), with 100,480,507 ratings. In Agarwal and Chen (2016, Sec.8.5), the authors illustrates an application of a version of RLFM to a large-scale problem with a Yahoo! Front Page dataset, whose training set includes "all pageviews by users with at least ten click ... and consist of 8 million users, approximately 4,300 items, and 1 billion binary observations." So, the model under consideration can in principle apply to, for instance, a few thousands stocks traded on the Tokyo Stock Exchange.

- *Shrinkage toward the regression means:* One interpretation of the model is that, for stocks with fewer observations, the expected value of y_{ij} tends to shrink to the mean determined by the regression terms (2)–(5), not to the origin. This feature can indeed handle the cold-start situations, shifting from the warm-start in a seamless manner.

3 Empirical analysis

3.1 Data

Dataset and preprocessing

In this empirical analysis, we use the FLEX Full (time resolution is micro second) kindly provided by Tokyo Stock Exchange (TSE), Japan Exchange Group. The data period ranges from January 4, 2019 to March 29, 2019 (58 business days). We divide the data into an estimation period from January 4 to February 28 (38 business days) and a prediction period from March 1 to March 29 (20 business days). Among those listed on the First Section of TSE, we focus on ninety-eight stocks out of the TOPIX 100 ingredients and ninety out of the Mid400 ingredients ($M = 188$).

First, we construct a limit-order book dataset (in micro seconds) updated every limit order, cancellation, and transaction for each business day and each stock from 9:00 to 15:00 except the lunch break 11:30–12:30. We divide the dataset into 60 sub datasets at 5 minute intervals (5×60 =300 minutes) excluding the closing records at 11:30 and 15:00. Then, we calculate the objective variables, ILOBS and realized volatility (denoted by RV hereafter), and the explanatory variables in the RLFM for each sub dataset. Consequently, we obtain 60 observations of the variables for each business day and each stock.

Variables

For stock i ($1 \leq i \leq M = 188$) and time bin j (5 minute intervals, $1 \leq j \leq N = 60$), we use the following variables.

- y_{ij}: liquidity/volatility measure:

(a) ILOBS in logarithms (logILOBS) calculated from the limit-order book up to 10 levels on both sides.

(b) RV in logarithms (logRV) calculated computed on tick-by-tick mid quote prices during the time bin.

- x_{ij}: logarithmic return (logRet.D), the number of transactions in logarithms (lognD), transaction volume (logDamt), imbalance between buy and sell transactions (ImbD), quoted spread (Spr), the number of quotes in logarithms (lognQ) and constant term ($s = 7$).
- w_i: Nikkei 225 Index (N225) dummy, TOPIX 17 series dummies (17 omitted), market capitalization (trillion JPY) in logarithms (logMCAP) on the previous day ($p = 18$).
  - TOPIX 17 series: 1. Foods, 2. Energy Resources, 3. Construction & Materials, 4. Raw Materials & Chemicals, 5. Pharmaceutical, 6. Automobiles & Transportation Equipment, 7. Steel & Nonferrous Metals, 8. Machinery, 9. Electric Appliances & Precision Instruments, 10. IT & Services, Others, 11. Electric Power & Gas, 12. Transportation & Logistics, 13. Commercial & Wholesale Trade, 14. Retail Trade, 15. Banks, 16. Financials (excluding Banks), 17. Real Estate.
- z_j: time bin polynomials $(j/N)^k$, $k = 1, 2, \ldots, 6$ ($q = 6$).

Summary statistics

First, we compute averages of the variables for each stock $i = 1, \ldots, M = 188$, over time bins $j = 1, \ldots, N = 60$, and over $T = 58$ days. Specifically, let $a_{ij,t}$ be a variable in y_{ij} or x_{ij} for stock i and time bin j on day t. Then, we compute its average as $\bar{a}_i = \sum_{j,t} a_{ij,t}/NT$ for each i and calculate summary statistics of $\bar{a}_i$ ($i = 1, \ldots, M = 188$).[8] Table 1 presents the summary statistics for each variable in y_{ij} and x_{ij} and for logMCAP in w_i. The stock attributes used for the dummy variables in w_i are summarized in Table 2.

Figure 1 shows the histogram of the averages $\bar{a}_i$ for each variable colored by N225 ingredients and Other stocks. We observe that the illiquidity (logILOBS), volatility (logRV), the number of transactions (lognD), and quoted spread (Spr) tend to be smaller for N225 ingredients, whereas the transaction volume (logDamt), the number of quotes (lognQ), and market capitalization (logMCAP) tend to be larger for them.

Second, we compute averages of the variables for each stock and each time bin, $\bar{a}_{ij} = \sum_t a_{ij,t}/T$, and then scale the averages as $\tilde{\bar{a}}_{ij} = (\bar{a}_{ij} - \bar{a}_i)/sd(a_i)$, where $sd(a_i)$ is the standard deviation over time bins and days. Figure 2 shows the boxplots of the scaled averages $\tilde{\bar{a}}_{ij}$ for each variable colored by N225 ingredients and Other. We observe notable intraday periodicities. The illiquidity (logILOBS) is high around the time the morning and afternoon sessions begin but decreases toward the end of the afternoon session. The number of transactions (lognD) and the quoted spread (Spr) show similar patterns. Such intraday variations are in line with that of the price impact (the coefficient of regressing price changes on order imbalances) observed in Cont et al. (2014). In addition, those variables for the N225 ingredients are higher than other stocks when they are high in the morning session and the beginning of the afternoon, but the relation is reversed in the rest of afternoon session. That is, we observe stronger intraday periodicities for N225 ingredients than Other.

On the other hand, the volatility (logRV), transaction volume (logDamt), and the number of quotes (lognQ) exhibit a roughly W- shaped pattern. They take the highest values in the first time bin, decline at

[8] For a variable in w_i, denoted by $a_{i,t}$, we compute its average as $\bar{a}_i = \sum_t a_{i,t}$ for each i and calculate their summary statistics.

Table 1: Summary statistics for variables in y_{ij}, x_{ij}, and w_i.

		Mean	SD	Min	Median	Max
y_{ij}	logILOBS	1.4713	1.7946	−6.3261	1.6740	5.6373
	logRV	−2.0052	0.3565	−2.7264	−2.0717	−0.7483
x_{ij}	logRet.D	0.0007	0.0033	−0.0087	0.0006	0.0139
	lognD	3.6260	2.6912	0.7955	2.6905	16.7226
	logDamt	7.4288	0.6147	6.1088	7.5542	8.8933
	ImbD	0.2610	2.8760	−7.5759	0.0749	13.5832
	Spr	8.2578	5.7738	1.7525	6.0097	33.4802
	lognQ	2.7772	0.4188	1.7550	2.8651	3.6946
w_i	logMCAP	−0.0602	1.2181	−2.3862	0.1018	3.0944

The summary statistics are obtained from 188 averages calculated by using all observations in the sample period for each stock.

a decreasing rate, slightly increase around the end of the morning session, rise after the afternoon session begins, decline at a decreasing rate, and then increase at an increasing rate until the end of the afternoon session. Ignoring the effect of the lunch break unique in TSE, such a pattern is in line with a U-shaped pattern of market activities observed in a number of previous studies.[9] Similarly to the logILOBS, lognD, and Spr, the logRV and logDamt for the N225 ingredients are larger (smaller) than Other stocks when they are large (small), whereas the relation is not so clear for lognQ especially in the afternoon session.

Third, we compute correlation coefficients among the variables for each stock over time bins and days. Table 3 summarizes them, where the means of 188 correlation coefficients are presented as lower triangular elements while the maximum values (in absolute value) are in upper triangular elements. Reflecting the similar intraday periodicity shown in Figure 2, the logILOBS is highly correlated wth Spr and lognD, whereas logRV is highly correlated with logDamt and lognQ. Some variables in x_{ij} are also highly correlated with each other. In particular, the correlation between lognD and Spr is quite strong with up to 0.99 for some stock. Considering the multicollinearity, we remove the intraday periodicities of the variables in x_{ij} prior to the estimation. The procedure of data processing is described in the following subsection.

3.2 Estimation and prediction procedures

Data processing

Prior to the estimation for each day, we process the data as follows:

1. Standardize x_{ij} using an average and a standard deviation over j for each i.

[9]For example, Wood et al. (1985) analyze NYSE-listed stocks and report a U-shaped pattern for minute-by-minute average returns and a reverse J-shaped pattern for the variability of returns. McInish and Wood (1992) report a crude J-shaped pattern for minute-by-minute spreads and Lee et al. (1993) report a U-shaped pattern for half-hour volumes and spreads. Additionally, Andersen and Bollerslev (1997) report a U-shaped pattern for five-minute absolute returns for S&P 500 stock index futures, although this drops and rises sharply before the close.

Figure 1: Histograms of averages for each variable colored by N225 ingredients and Other stocks.

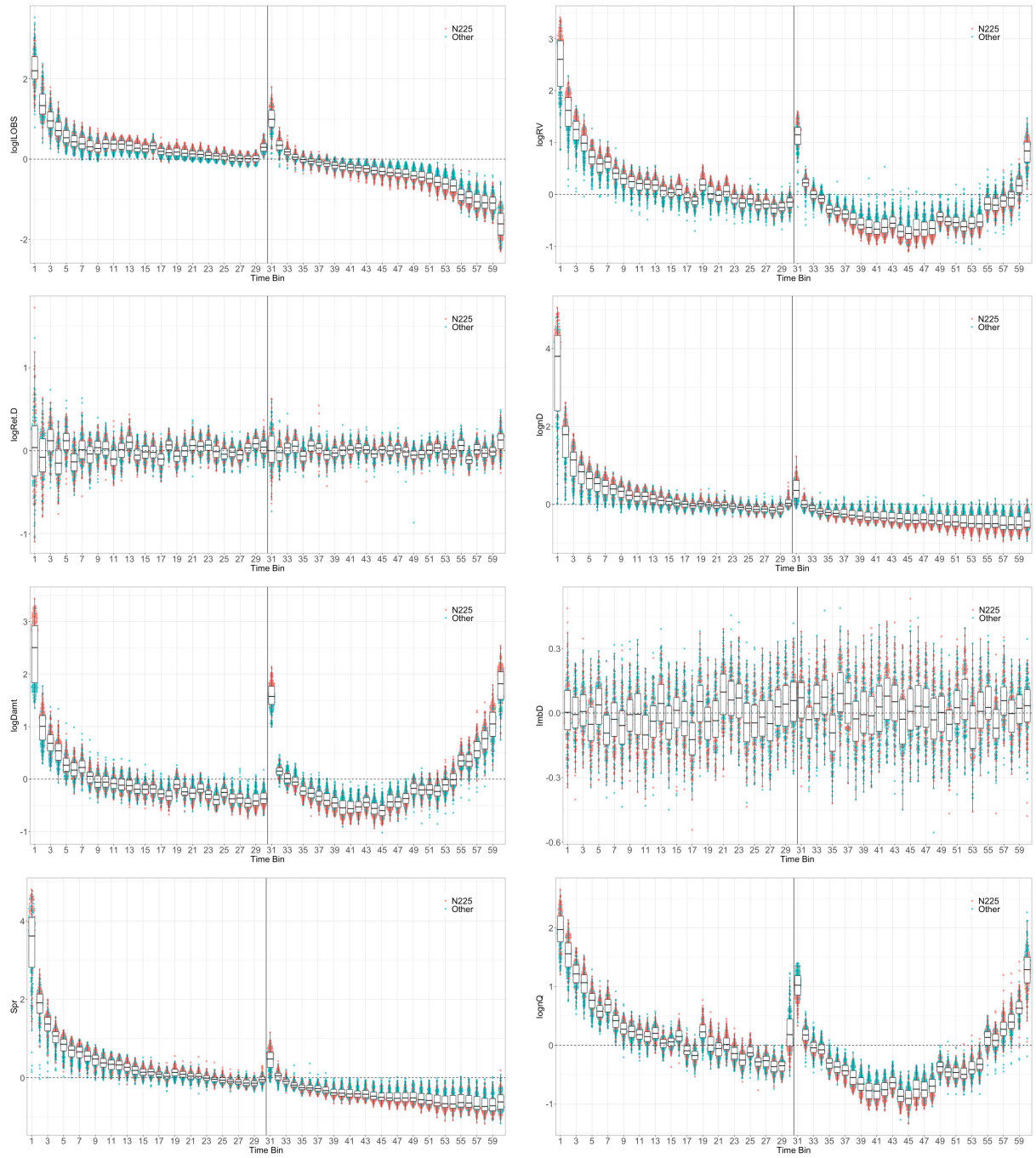

Figure 2: Boxplots of scaled averages for each variable and each time bin $j = 1, \ldots, 60$. The vertical line separates the morning and afternoon sessions.

Table 2: Summary of stock attributes for w_i.

	TOPIX 17 series																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	Total
N225	4	2	2	8	6	9	4	4	9	6	4	5	5	3	5	7	3	86
	2.1	1.1	1.1	4.3	3.2	4.8	2.1	2.1	4.8	3.2	2.1	2.7	2.7	1.6	2.7	3.7	1.6	45.7
Other	5	0	6	11	4	4	1	6	10	19	2	6	5	10	7	5	1	102
	2.7	0.0	3.2	5.9	2.1	2.1	0.5	3.2	5.3	10.1	1.1	3.2	2.7	5.3	3.7	2.7	0.5	54.3
Total	9	2	8	19	10	13	5	10	19	25	6	11	10	13	12	12	4	188
	4.8	1.1	4.3	10.1	5.3	6.9	2.7	5.3	10.1	13.3	3.2	5.9	5.3	6.9	6.4	6.4	2.1	100

For each group of N225 ingredients, Other stocks, and in total (Total), the top row presents the number of stocks for each industry (1, 2, ..., 17) and in total (Total), whereas the bottom shows the proportion (%) to $M = 188$ stocks.

Table 3: Summary of correlation coefficients among variables in y_{ij} and x_{ij}.

		y_{ij}		x_{ij}					
		logILOBS	logRV	logRet.D	lognD	logDamt	ImbD	Spr	lognQ
y_{ij}	logILOBS	–	0.78	−0.14	0.79	0.40	−0.12	0.84	0.58
	logRV	0.40	–	0.13	0.76	0.86	−0.15	0.77	0.94
x_{ij}	logRet.D	−0.01	0.00	–	0.28	−0.13	0.61	0.28	0.12
	lognD	0.45	0.48	−0.01	–	0.51	−0.11	0.99	0.67
	logDamt	0.05	0.69	0.00	0.22	–	−0.18	0.49	0.90
	ImbD	−0.01	−0.00	0.45	−0.01	−0.00	–	−0.10	−0.18
	Spr	0.54	0.54	−0.00	0.82	0.18	−0.01	–	0.65
	lognQ	0.25	0.81	0.00	0.30	0.75	−0.01	0.34	–

The lower and upper triangular elements present means and maximum values (in absolute value) of 188 correlation coefficients calculated by using all observations in the sample period for each stock, respectively.

2. Replace missing or infinite values of y_{ij} (we observe $-\infty$ for logRV when RV is zero) with an average over j for each i.[10]

3. Replace missing values of x_{ij} with zero.

4. Remove intraday periodicities in x_{ij} by regressing stacked vectors of x_{ij} for each variable on z_j.[11]

Figure 3 shows the boxplots of fitted values of the regressions of 58 days for each variable. We observe that the intraday periodicity shown in Figure 2 are captured well. In addition, Table 4 summarizes the correlation coefficients among the variables after removing the periodicities, where the means of 58 correlation coefficients are presented as lower triangular elements while the maximum values (in absolute value) are in

[10] This is merely for convenience and is not necessarily the best way to treat them. There is room for exploring a better way but we do not pursue it here because it is not the main focus of the paper.

[11] Specifically, let $a_{ij,t}$ be a variable in x_{ij} for stock i and $a_{i,t} = (a_{i1,t}, a_{i2,t}, \ldots, a_{iNt})'$. Then, we regress the $(MN \times 1)$-vector, $a_t = (a_{1,t}, a_{2,t}, \ldots, a_{M,t})'$, on the $(MN \times (q+1))$-matrix, Z_j which is constructed by stacking a $(N \times 1)$-vector of ones and the $(N \times q)$-matrix of time bin polynomials up to degree $q = 6$, z_j, for M times. We implement the regression for each variable in x_{ij} and each day t and replace x_{ij} with the standardized residuals.

upper triangular elements. We confirm that all of the maximum correlation coefficients (in absolute value) are less than 0.80 which is the most typical cutoff for multicollinearity mentioned in Berry and Feldman (1985).

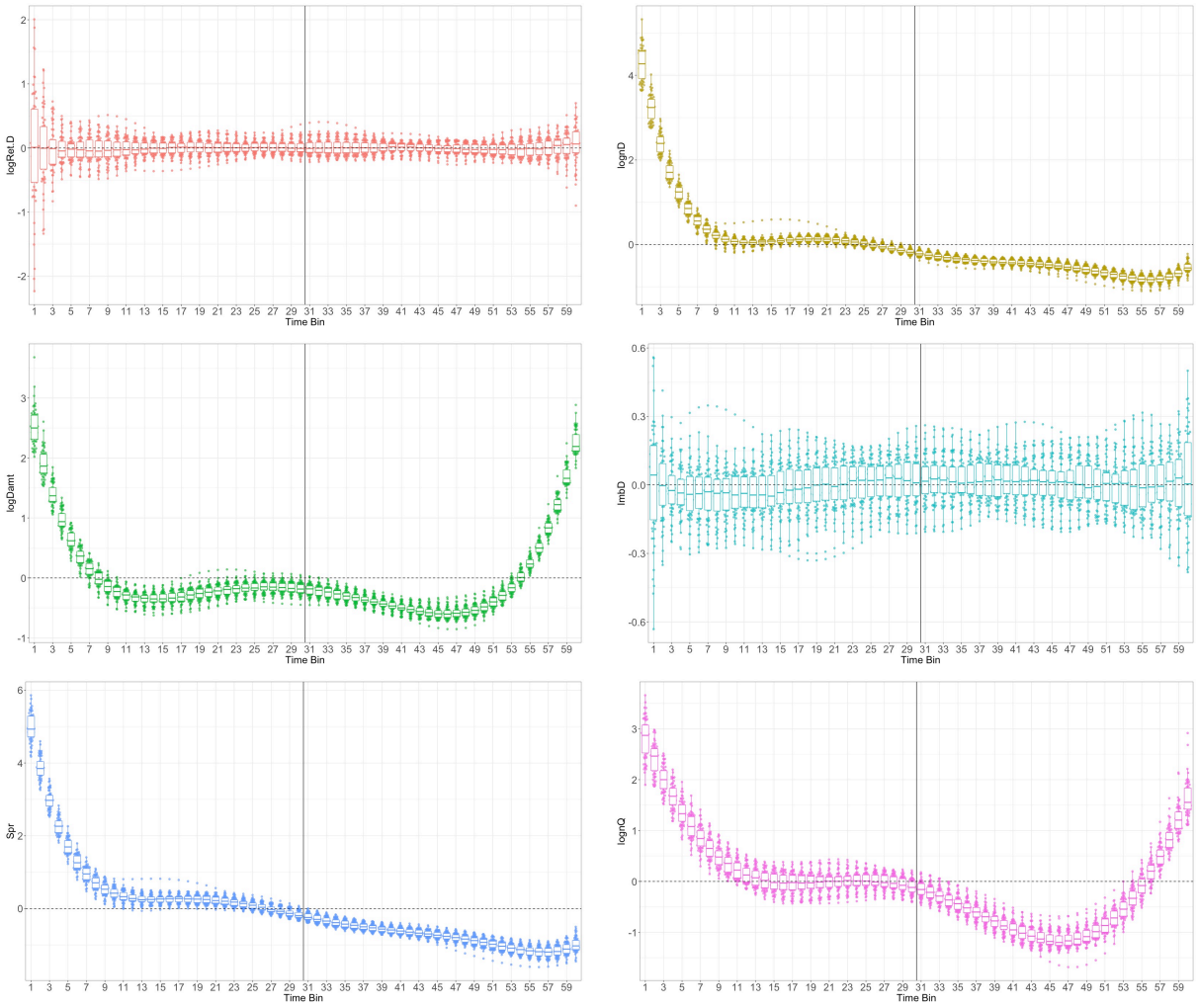

Figure 3: Boxplots of fitted values of 58 regressions on time bin polynomials for each variable and each time bin $j = 1, \ldots, 60$. The vertical line separates the morning and afternoon sessions.

Estimation

Adopting the Monte Carlo EM (MCEM) algorithm proposed by Agarwal and Chen (2016), we estimate the model parameter $\Theta = (b, g_0, d_0, G, D, \sigma_\alpha, \sigma_\beta, \sigma_u, \sigma_v)$ and the (means of) latent factors $\Delta = (\alpha_i, \beta_j, u_i, v_j)$ for each day of the estimation period. In the MCEM algorithm, E-step is implemented using samples generated by Gibbs sampler. Following Agarwal and Chen (2016), we iterate 20 EM computations using 100 Gibbs samples (drawn after 10 burn-in samples).

Table 4: Summary of correlation coefficients among variables in x_{ij} after removing the periodicities.

	logRet.D	lognD	logDamt	ImbD	Spr	lognQ
logRet.D	–	-0.08	0.12	0.52	0.08	0.12
lognD	-0.01	–	0.19	-0.04	0.66	0.20
logDamt	0.00	0.14	–	0.10	0.17	0.77
ImbD	0.48	-0.01	0.00	–	0.04	0.08
Spr	0.00	0.58	0.12	0.00	–	0.25
lognQ	0.00	0.13	0.72	0.00	0.17	–

The lower and upper triangular elements present means and maximum values (in absolute value) of 58 correlation coefficients calculated by using $M \times N = 188 \times 60 = 11,280$ observations for each day of estimation period, respectively.

Prediction[12]

The prediction procedure is as follows:

1. Extract $p\,(= 20, 50, 80)\,\%$ of prediction target pairs (i, j) randomly from the stock-time matrix (y_{ij}).[13]
2. Standardize x_{ij} and replace missing or infinite values of y_{ij} and x_{ij} as specified in the data processing procedure 1-3.
3. Replace the corresponding y_h and x_h $(h = 1, \ldots, H)$ with the average of y_{ij} over j for each i and 0, respectively.
4. Remove the intraday periodicities of x_{ij} as specified in the data processing procedure 4.
5. Estimate the RLFM for each day of the prediction period as described above.
6. Given the estimated parameter values and latent factors, compute a predicted valued of y_h, denoted by $\hat{y}_h$, using actual values of x_h.[14]
7. Calculate the root mean squared error (RMSE) and mean absolute error (MAE):

$$RMSE = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(y_h - \hat{y}_h)^2}, \qquad MAE = \frac{1}{H}\sum_{h=1}^{H}|y_h - \hat{y}_h|$$

For model comparison, we estimate a linear regression model with regressors x, w, and z (LM0), LM0 with fixed effects of stocks and time bins (LMD), and LM0 with random effects of stocks and time bins (LMM), as well as the user- and item-based collaborative filtering (UBCF and IBCF) and the singular value decomposition (SVD).

All analyses are done by the statistical software `R`. We estimate the LMM using `lmer` function included in '`lme4`' package and the UBCF, IBCF, and SVD by using `Recommender` function included in '`recommenderlab`' package with parameter `method` specified as "UBCF", "IBCF", and "funkSVD", respectively.

[12] In this paper, the prediction means interpolation where we predict the missing target variable by using explanatory variables observed at the same time. In the case of extrapolation where we predict the target variable in the future, we may use explanatory variables on the previous day or the ones averaged over the last few days. We leave this for future research.

[13] We exclude the missing or infinite observations of y_{ij} from the prediction target pairs.

[14] Intraday periodicities of x_h are removed by subtracting the fitted values obtained in the prediction procedure 4.

3.3 Results

(a) logILOBS

Figure 4 summarizes parameter estimates of the RLFM with $r = 3$.[15] Because the objective variable is the illiquidity measure, ILOBS in logarithms, high (low) values indicate low (high) liquidity. Among the estimates of b, the number of transactions (lognD), quoted spread (Spr), and the number of quotes (lognQ) are significantly positive (liquidity deteriorates when they increase) whereas the transaction volume (logDamt) is significantly negative (liquidity improves when they decrease).

For the estimates of g_0 (stock attributes), we observe different effects among TOPIX 17 series (transformed to satisfy the sum to zero constraint for each of 38 estimation days[16]). In particular, series #5 (Pharmaceutical), #8 (Machinery), and #12 (Transportation & Logistics), are significantly positive whereas #2 (Energy Resources), #15 (Banks), and #16 (Financials excluding Banks) are significantly negative. Further, the N225 dummy and the market capitalization (logMCAP) are significantly negative and positive, respectively (liquidity is higher for the N225 ingredients but is lower for the large stocks). For the estimates of d_0 (time attributes), all time bin polynomials are significant, suggesting nonlinear intraday periodicity in the logILOBS.

Figure 5 summarizes the estimates of latent factors α_i and β_j. We observe that N225 ingredients tend to be liquid as expected from the estimates of g_0. We also observe an intraday periodicity in the estimates of β_j (time effect) as suggested from the estimates of d_0. In the morning session, the liquidity is relatively high right after the opening and then gradually deteriorates, marginally improves around the middle, and deteriorates again toward the lunch break. In the afternoon session, on the other hand, it reaches the worst right after the opening and then gradually improves at an increasing rate toward the closing.

Figure 6 summarizes the model fit comparisons.[17] There are no significant difference among the RLFM with $r = 3$ and 7. Our approach (RLFM) is slightly inferior to other regression models. Specifically, the LMD and LMM shows equally well performance in both RMSE and AIC and the RLFM follows with a narrow margin.

Figure 7 summarizes the predictive performance.[18] Our approach (RLFM) is comparable to the LMM, UBCF, and SVD for the middle missing rate p (50%) and slightly inferior to some of them for other p's. Specifically, the SVD ($r = 3$) and UBCF perform well and the RLFM follows for small p (20%) whereas the LMM does well and the RLFM follows with a narrow margin for large p (80%). These results show that our approach is *not* uniformly optimal among the competing models, nevertheless they suggest that it is robust to the change of p and is therefore capable of dealing with both the rich and scarce data reasonably well.

(b) logRV

Figure 8 summarizes parameter estimates of the RLFM with $r = 3$.[19] Among the estimates of b, the transaction volume (logDamt), quoted spread (Spr), and the number of quotes (lognQ) are significantly

[15] We also estimate the RLFM with $r = 7$. The results are qualitatively the same and thus are omitted.

[16] Specifically, let $\theta = (a_1, a_2, \ldots, a_{16}, a_{17})'$ be a vector of parameter estimates on the industry dummies, where $a_1, a_2, \ldots, a_{16}$ are estimated coefficients on 16 dummies (17 is omitted) and a_{17} is the estimate on a constant term. Then, we transform θ to $\tilde{\theta} = \theta - \bar{\theta}$, where $\bar{\theta} = \sum k = 1^{17} a_k/17$.

[17] Because the LM0 is significantly worse than other models, its results are omitted.

[18] Because the LM0 and LMD are significantly worse than other models and other SVD with $r = 2, \ldots, 15$ perform similarly to those with $r = 3$ and 7, their results are omitted for brevity.

[19] We also estimate the RLFM with $r = 7$. The results are qualitatively the same and thus are omitted.

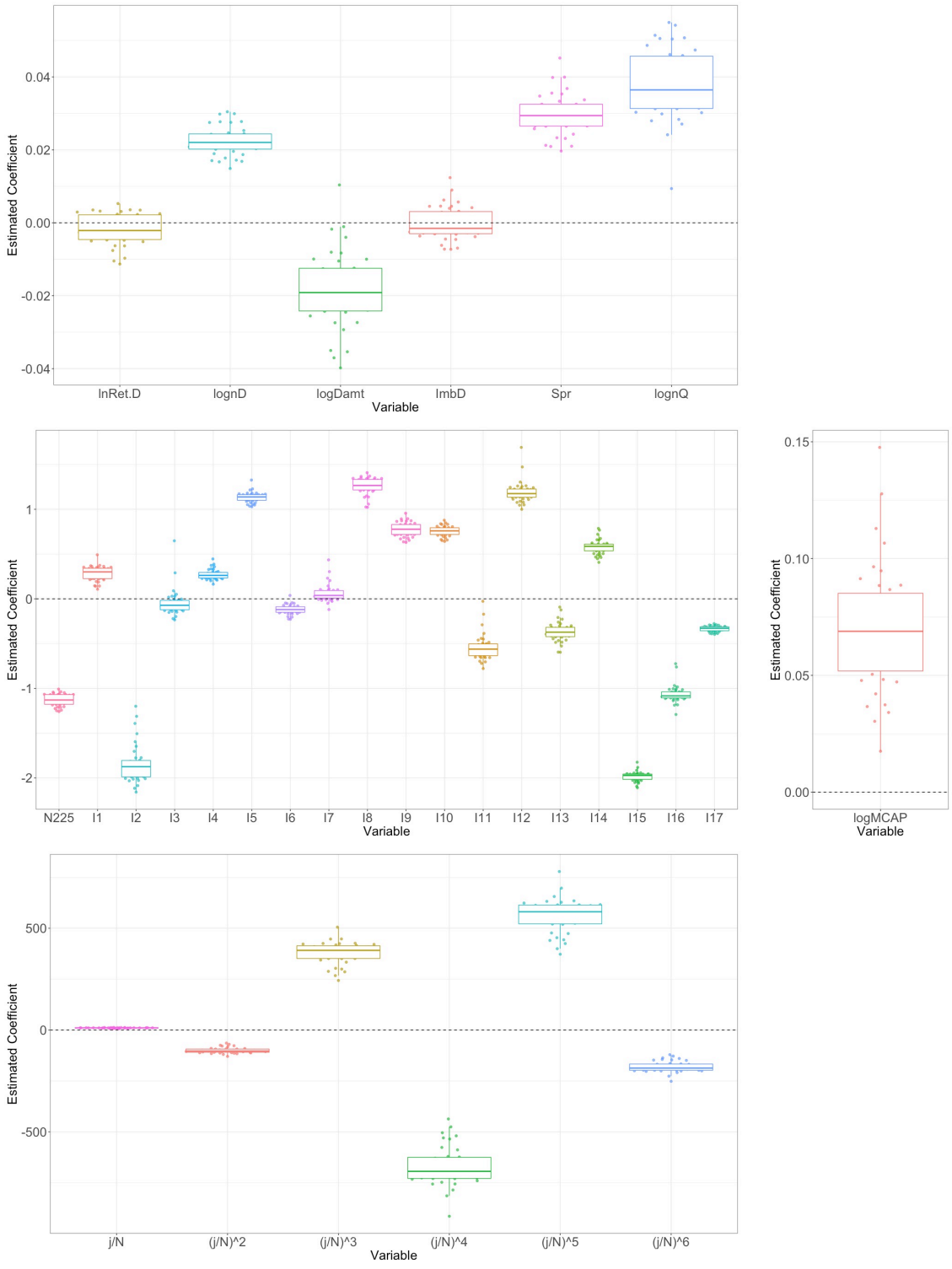

Figure 4: Parameter estimates of RLFM ($r = 3$) for logILOBS. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Top panel shows the coefficients b on market activities of individual stocks x_{ij}, middle g_0 on static attributes of individual stocks w_i, and bottom d_0 on time bin polynomials z_j. The estimates of g_0 on TOPIX 17 series dummies (I1, I2, ..., I17) are transformed to satisfy the sum to zero constraint for each of 38 estimation days (see footnote 16 for the details).

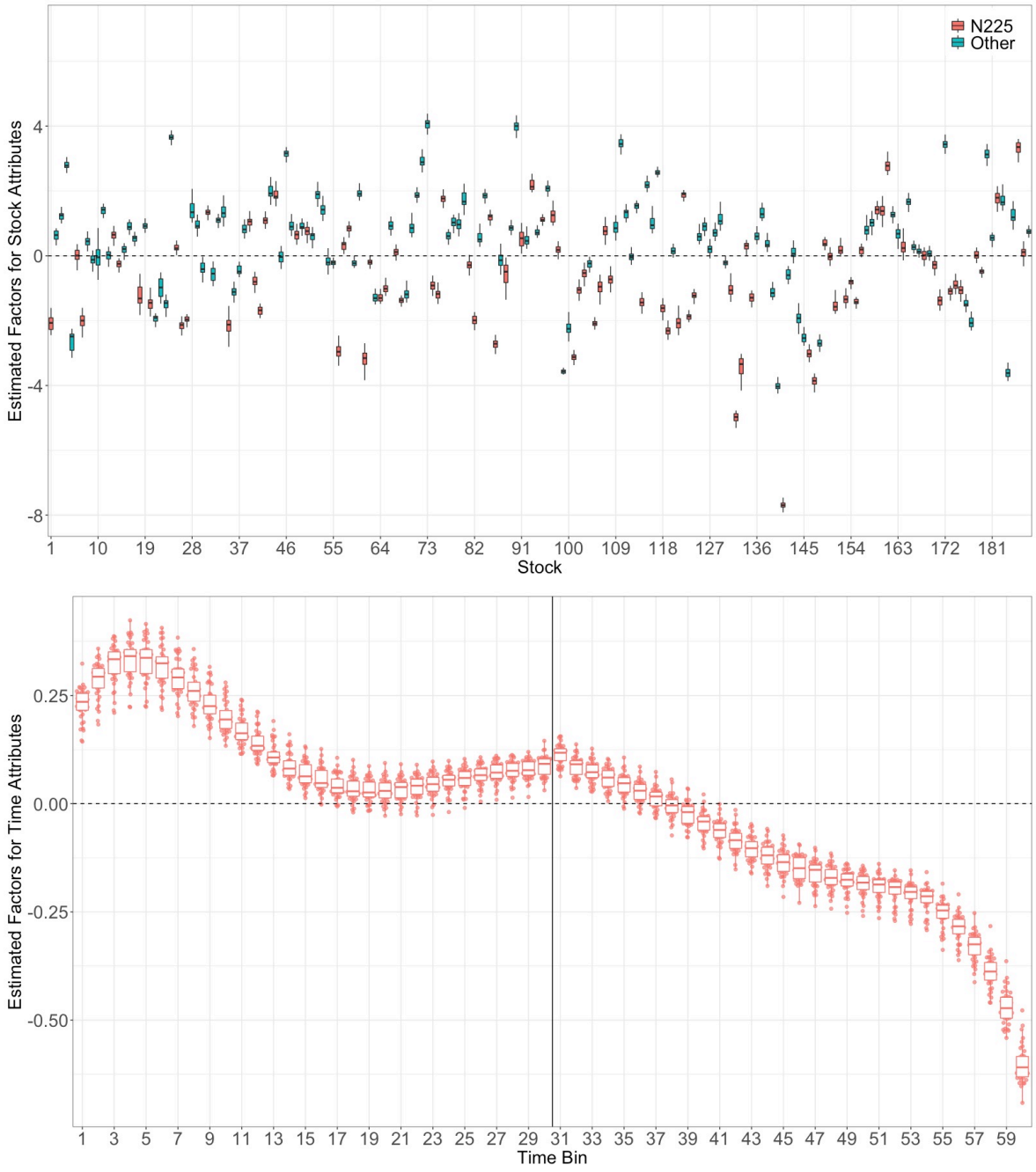

Figure 5: Latent factors of RLFM ($r = 3$) for logILOBS. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Top panel shows the factors of stock attributes α_i, $i = 1, 2, \ldots, M = 188$, colored by N225 ingredients and Other stocks, whereas the bottom the factors of time attributes β_j with the vertical line separating the morning and afternoon sessions.

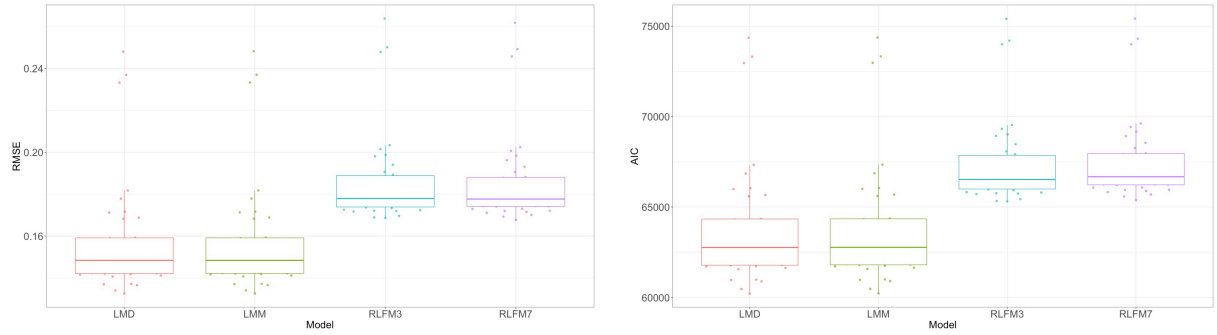

Figure 6: Model fit for logILOBS. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Left and right panels show the RMSE and AIC, respectively.

Figure 7: Predictive performance for logILOBS. Boxplots are constructed from the prediction results for 20 business days from March 1, 2019, to March 29, 2019. Left and right panels show the RMSE and MAE, respectively, for missing rates p equal to 20% (top), 50% (middle), and 80% (bottom).

positive, and so is the number of transactions (lognD) to a lesser extent (volatility increases when they increase).

For the estimates of g_0 (stock attributes), we observe mostly marginal effects among TOPIX 17 series except a few industries. series #2 (Energy Resources) and #17 (Financial excluding Banks). In particular, series #2 (Energy Resources) is significantly positive whereas #12 (Transportation & Logistics) and #17 (Real Estate) are significantly negative. The N225 dummy is significantly negative (volatility is lower for the N225 ingredients) as well. On the other hand, the effect of the market capitalization (logMCAP) is indeteminate. For the estimates of d_0 (time attributes), similarly to logILOBS, all time bin polynomials are significant, suggesting nonlinear intraday periodicity in the volatility.

Figure 9 summarizes the estimates of latent factors α_i and β_j. We observe that N225 ingredients tend to be less volatile as expected from the estimates of g_0. We also observe an intraday periodicity in the estimates of β_j (time effect) as suggested from the estimates of d_0. In the morning session, the volatility is substantially high right after the opening and then gradually decreases toward the lunch break. In the afternoon session, the volatility gradually decreases over time, reaches the lowest in time bin 48 (13:55–14:00) and increases toward the closing.

Figure 10 summarizes the model fit comparisons.[20] There are no significant difference among our approach (RLFM) and other regression models (LMD and LMM).

Figure 11 summarizes the predictive performance.[21] Our approach (RLFM) outperforms the other models for small and middle missing rates (p = 20% and 50%) and comparable to the LMM for large one (p = 80%). Compared to the results for logILOBS, the RLFM performs quite better than the LMM and the other models. Again, these results suggest that our approach is capable of handling both the rich and scarce data reasonably well though it does *not* always perform best.

In practical situations, one would not know *a priori* the precise degree of occurrences of "no liquidity" situations resulting in missing observations of the market quality measures considered here. Since our approach is robust to the variability of scarcity of observations, it can alleviate the burden of market participants from selecting an "optimal" model among alternatives.

(c) logILOBS vs logRV

Figure 12 shows two dimensional box plots on the estimates of b and g_0 illustrated in (a) and (b) above. Among the estimates of b, the number of transactions (lognD), quoted spread (Spr), and the number of quotes (lognQ) are significantly positive in both dimensions. This indicates that the market tend to be less liquid and more volatile when there are more transactions or quotes with wider spread. On the other hand, the transaction volume (logDamt) is significantly negative in logILOBS but positive in logRV, suggesting that the market tend to be more liquid and more volatile when there are larger transactions. This is probably because large transactions, which may cause price change and hence increase mid-quote RV, occur when there are enough quotes on the limit order book, i.e., the market is liquid, resulting in higher volatility with market more liquid in the same 5-minute interval.

Among the estimates of g_0, we observe diverse effects. For example, the N225 dummy is significantly negative in both dimensions, indicating that the N225 ingredients tend to be more liquid and less volatile. In addition, the market capitalization (logMCAP) is significantly positive in logILOBS whereas its effect is

[20] Because the LM0 is significantly worse than other models, its results are omitted.

[21] Because the LM0 and LMD are significantly worse than other models and other SVD with $r = 2, \ldots, 15$ perform similarly to those with r = 3 and 7, their results are omitted for brevity.

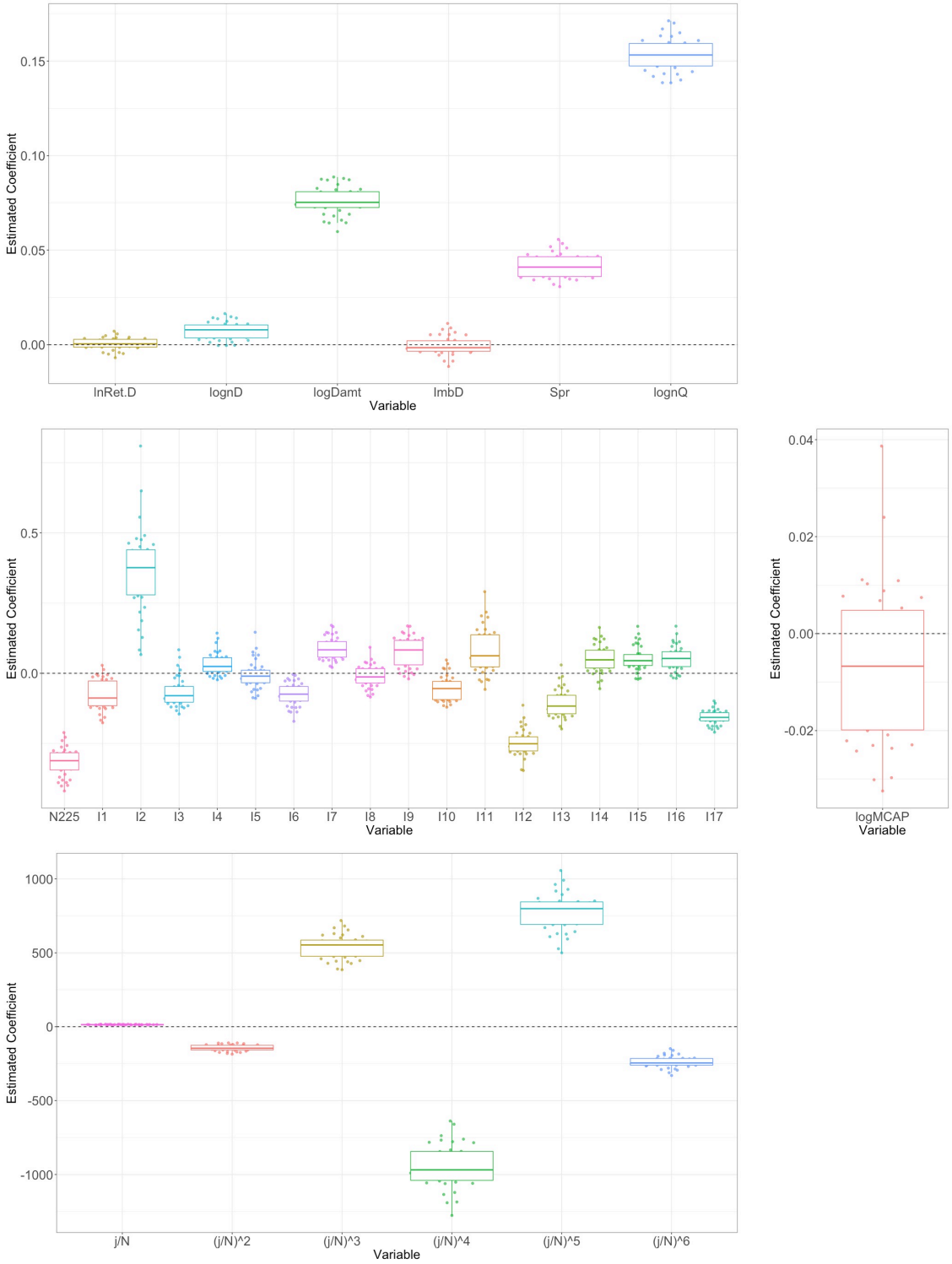

Figure 8: Parameter estimates of RLFM ($r = 3$) for logRV. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Top panel shows the coefficients b on market activities of individual stocks x_{ij}, middle g_0 on static attributes of individual stocks w_i, and bottom d_0 on time bin polynomials z_j. The estimates of g_0 on TOPIX 17 series dummies (I1, I2, ..., I17) are transformed to satisfy the sum to zero constraint for each of 38 estimation days (see footnote 16 for the details).

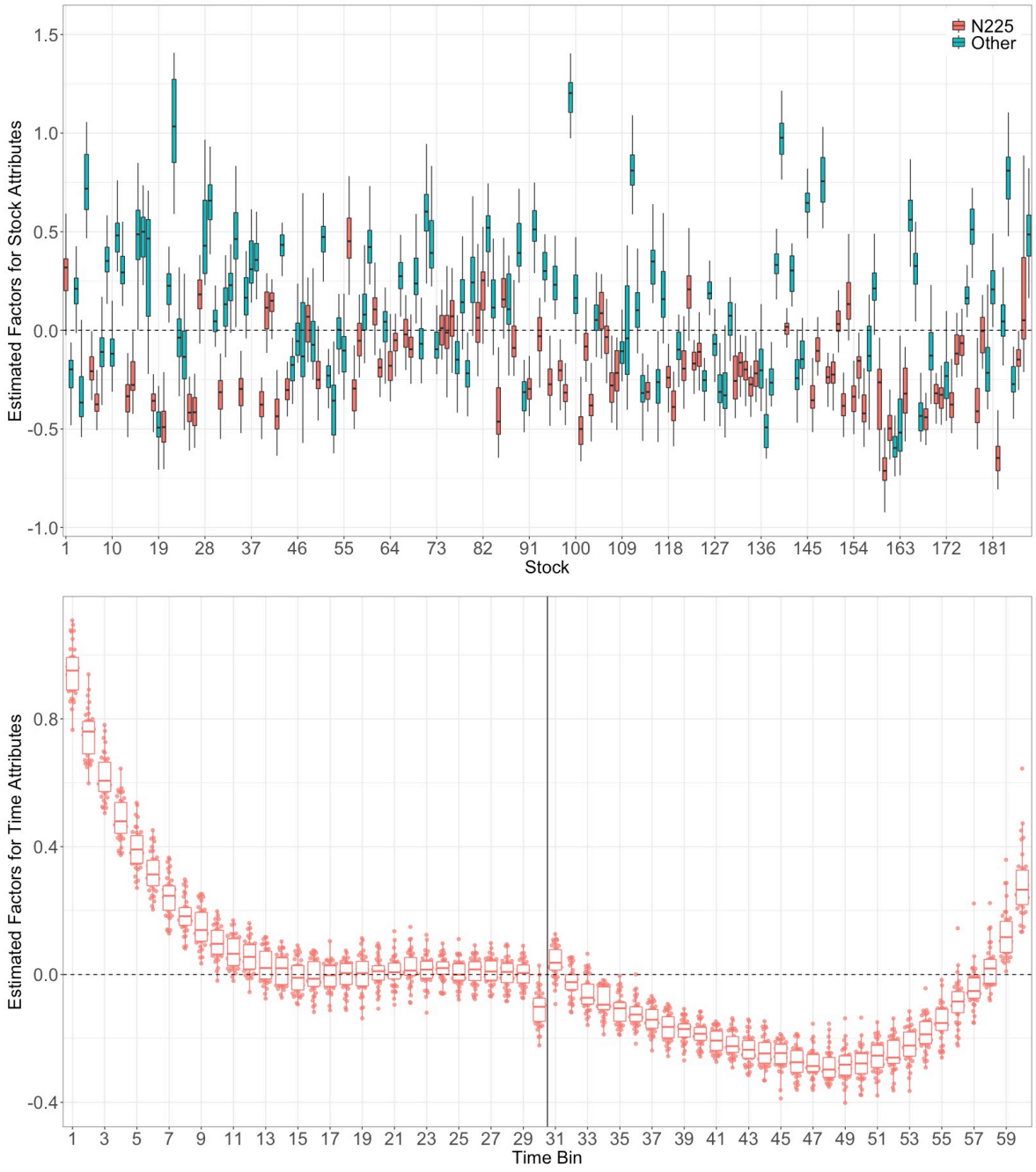

Figure 9: Latent factors of RLFM ($r = 3$) for logRV. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Top panel shows the factors of stock attributes α_i, $i = 1, 2, \ldots, M = 188$, colored by N225 ingredients and Other stocks, whereas the bottom the factors of time attributes β_j with the vertical line separating the morning and afternoon sessions.

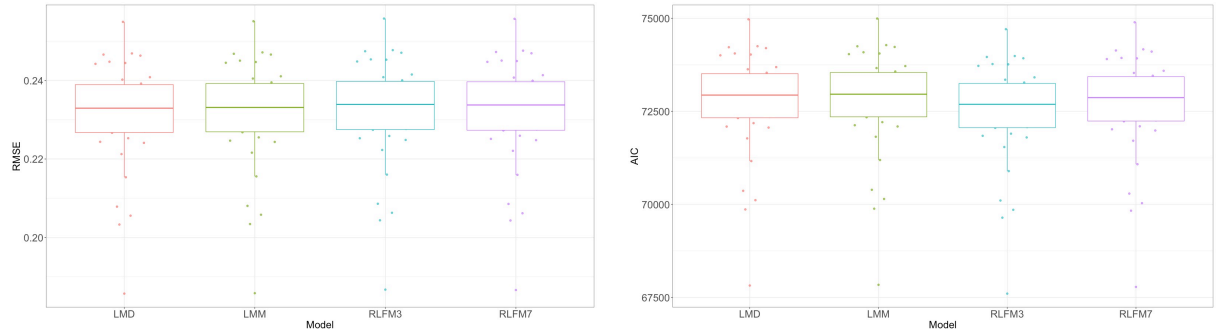

Figure 10: Model fit for logRV. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. Left and right panels show the RMSE and AIC, respectively.

Figure 11: Predictive performance for logRV. Boxplots are constructed from the prediction results for 20 business days from March 1, 2019, to March 29, 2019. Left and right panels show the RMSE and MAE, respectively, for missing rates p equal to 20% (top), 50% (middle), and 80% (bottom).

indeterminate in logRV. These results suggest that the N225 ingredients are more liquid and less volatile but become less liquid for stocks with larger logMCAP.

We also observe diverse industry effects. In particular, series #13 (Commercial & Wholesale Trade) and #17 (Real Estate) are significantly negative in both dimension, indicating that stocks of these industries tend to be more liquid and less volatile. On the other hand, several industries shows the opposite effects in liquidity and volatility. For example, series #2 (Energy Resources) is significantly negative in logILOBS but positive in logRV whereas #12 (Transportation & Logistics) is significantly positive in logILOBS but significantly negative in logRV.

This time we fit the model separately to each of logILOBS and logRV, and paired the corresponding estimated parameters to better understand the fitted results, as shown. However, the framework under investigation is broad enough to treat the two, or even more, measures jointly into a *single* model. This can be done by defining a new response variable as a *tensor* of mode 3, each element of which is denoted as y_{ijk}, representing the asset i's j-th measure at time bin k , $j = 1, 2$. In this extended model, the dependency between logILOBS and logRV can explicitly be incorporated.

(d) Illustrative examples

To illustrate how the models work in predicting y_{ij}, we present the predicted values for some stocks on some days. Based on the predictive performance, measured by RMSE, of the RLFM for each stock and each day, we selected a pair of Japan Tobacco Inc. and Alfresa Holdings Corporation for logILOBS, and another pair of Kyowa Exeo Corp. and SBI Holdings Inc. for logRV. The stock attributes are summarized in Table 5.

Figure 13 shows the predicted values, with observed (after data processing) and target (treated as missing and replaced with the daily average in prediction) values, for logILOBS of Japan Tobacco and Alfresa Holdings on March 25, 2019. The predictive performance for Japan Tobacco on that day is bad irrespective of models and missing rates. In particular, for $p = 80\%$, the RLFM performs the worst for Japan Tobacco whereas it performs the best for Alfresa on the day. For Japan Tobacco, we observe that all the models failed to capture the plummets and plunges in the afternoon session even for small missing rate $p = 20\%$. For Alfresa, on the other hand, the logILOBS is less volatile and moves around the daily average. These results suggest that the models may not perform well when the target variable (logILOBS here) varies substantially possibly due to unexpected/accidental market activities. In addition, the predicted values tend to shrink toward the daily average as the missing rate increases from $p = 20\%$ to 80% because more target values are replaced with the average in the data processing.

Figure 14 shows the similar ones for logRV of Kyowa Exeo and SBI Holdings on March 11, 2019. The predictive performance for Kyowa Exeo on that day is bad irrespective of models and missing rates. In particular, for $p = 80\%$, the RLFM performs the second-worst for Kyowa Exeo whereas it performs the best for SBI Holdings on the day.[22] We observe that the models failed to capture the large fluctuations for Kyowa Exeo. The results also suggest that they may not perform well when the target variable (logRV here) varies substantially.

[22] We did not select the stock and day for which the RLFM performs the worst because there are only four observations for the stock on the day.

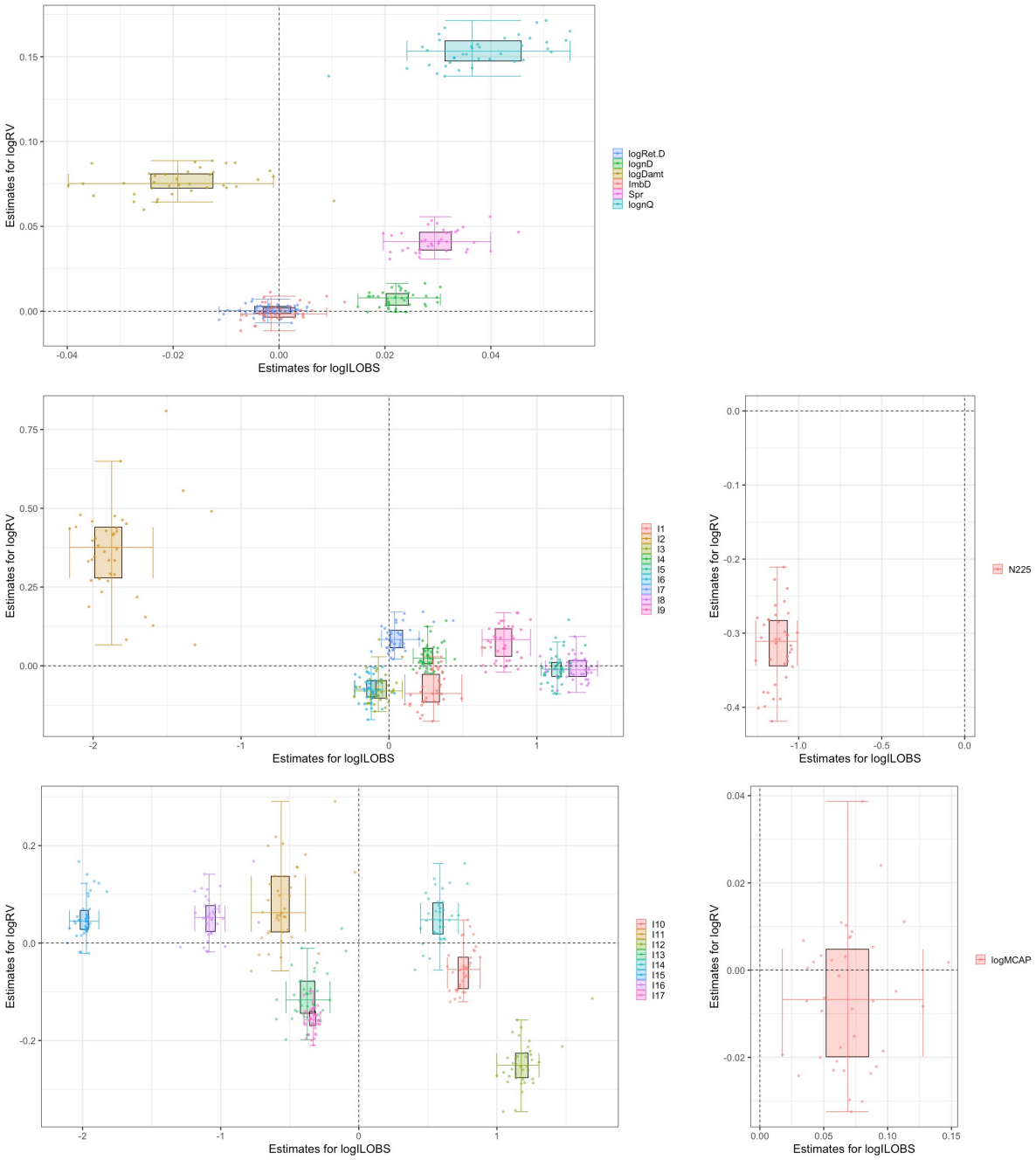

Figure 12: Parameter estimates of RLFM ($r = 3$) for logILOBS and logRV. Boxplots are constructed from the estimation results for 38 business days from January 4, 2019, to February 28, 2019. The horizontal and vertical axes correspond to the estimates for logILOBS and logRV, respectively. Top panel shows the estimates of b, middle g_0 on TOPIX 17 series #1–#9 (left) and N225 dummy (right), and bottom g_0 on the series #10–#17 (left) and logMCAP (right). The estimates of g_0 on TOPIX 17 series dummies (I1, I2, …, I17) are transformed to satisfy the sum to zero constraint for each of 38 estimation days (see footnote 16 for the details).

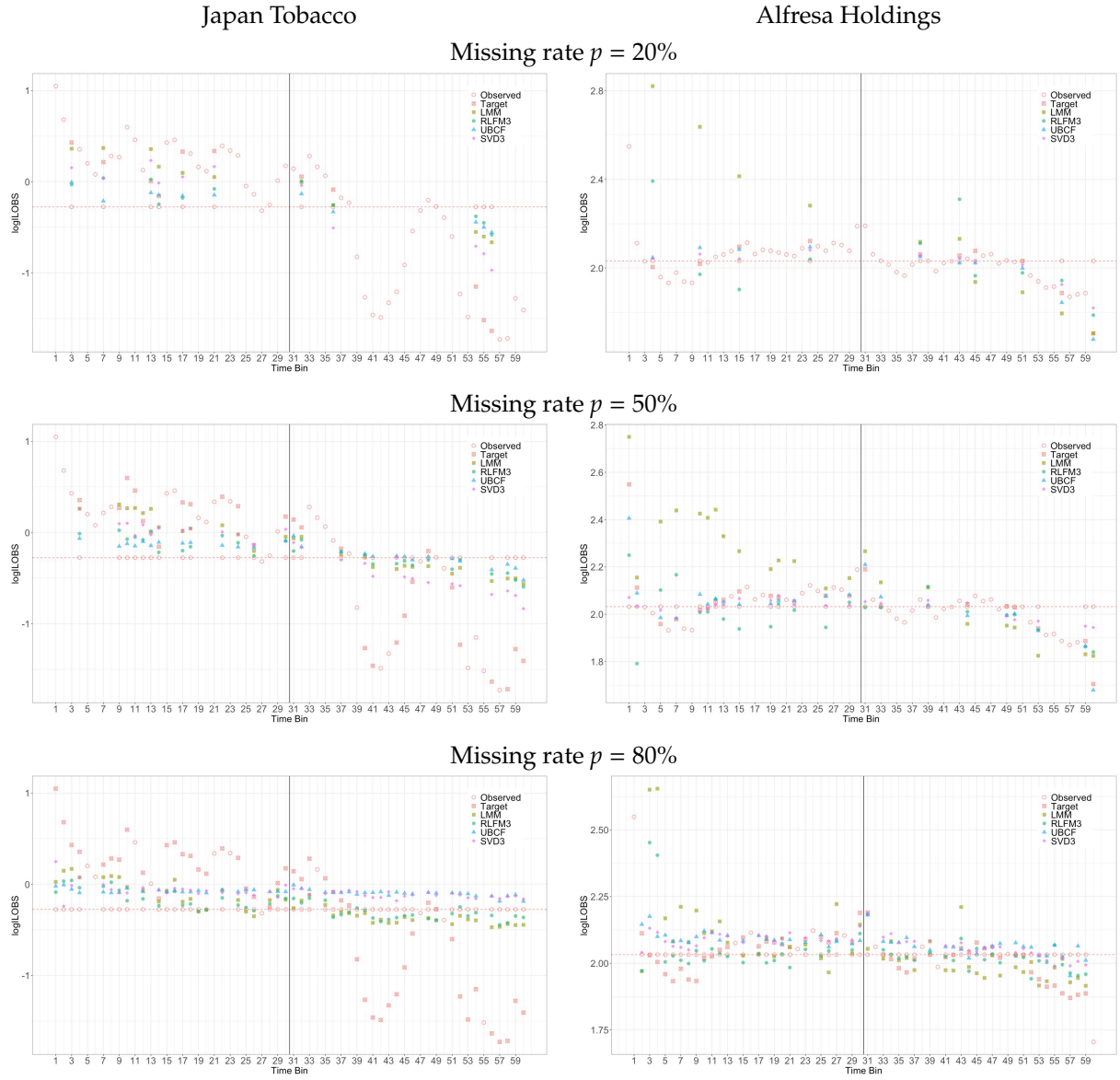

Figure 13: Predicted values for logILOBS of Japan Tobacco Inc. (left) and Alfresa Holdings Corporation (right) on March 25, 2019, with observed (after data processing) and target (treated as missing and replaced with the daily average, indicated as the dashed line, in prediction) values. The vertical line separates the morning and afternoon sessions.

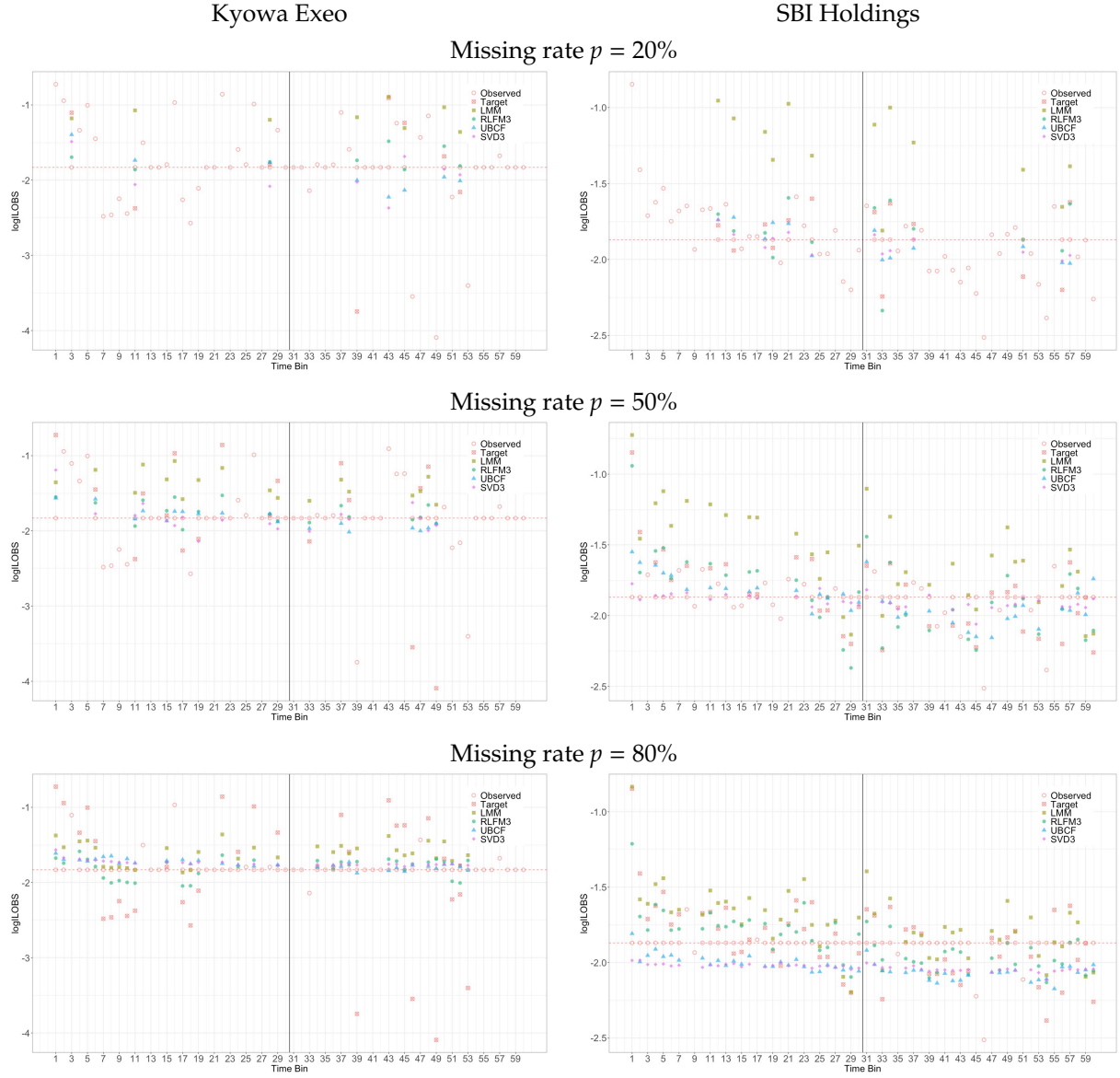

Figure 14: Predicted values for logRV of Kyowa Exeo Corp. (left) and SBI Holdings Inc. (right) on March 11, 2019, with observed (after data processing) and target (treated as missing and replaced with the daily average, indicated as the dashed line, in prediction) values. The horizontal dashed line indicates the average of logRV over the day. The vertical line separates the morning and afternoon sessions.

Table 5: Attributes of selected stocks.

Stock	N225	TOPIX 17 series		MCAP	Rank
Japan Tobacco	Yes	1.	Foods	5.5787	11th
Alfresa Hodings	No	13.	Commercial & Wholesale Trade	0.7630	107th
Kyowa Exeo	No	3.	Construction & Materials	0.3534	138th
SBI Holdings	No	16.	Financials (excluding Banks)	0.5716	118th

N225 indicates whether it is a N225 ingredient or not. MCAP and Rank present a value of market capitalization (trillion JPY) averaged over 20 business days from from March 1, 2019, to March 29, 2019, and its rank (in descending order) in 188 stocks considered, respectively.

3.4 Discussion

Robustness to the data processing. In the empirical analyses above, we considered a possible multicollinearity and used the variables after removing the intraday periodicities. To check if this data processing affects the main results, we conducted the same analyses using the variables without removing the periodicities. For the logILOBS, the results were qualitatively the same. For the logRV, on the other hand, the results for the RLFM were qualitatively the same again, whereas the predictive performance of the LMM was considerably better than the one using the periodicity-adjusted data. That is, our framework was robust to the data processing but the LMM was not. In addition, it is worth mentioning that the intraday periodicities captured by the estimated latent factor β_j, shown in Figures 5 and 9, were weaker when we used the data without removing the periodicities. The details are available upon request.

Robustness to the data period. We also checked if the data period affects the main results by using the data for the last three months of the year 2018.[23] We obtained qualitatively the same results and confirmed that our methodology worked fine in a different data period. Again, the details are available upon request.

Motivation and potential applications of the model. A natural question may arise about the motivation of "estimating" the "true" liquidity of an illiquid stock when there is indeed "no depth," or no single order placed on the limit-order book. Market participants usually see such situations as "no liquidity"; there might then be no use for her to assess a "theoretical" value because she can never initiate a trade against nothing on the book. We argue that computing theoretical values for liquidity can be useful especially when it is carried out in a prediction context. That is, after the closure of the market on the preceding day, by predicting the intraday liquidity of all the stocks in the trading universe for the following day, she can simulate and develop more realistic and feasible, finer trading strategies along with the time axis. In the meantime, volatility prediction for all the stocks in the universe shall certainly be of practical importance. Considering of covariances/correlations between pairs of the stocks is in principle possible in an extended framework, for instance, by introducing a new axis to represent another asset against which covariance/correlation is to be computed and by defining a new response variable as a tensor of mode 3, each element of which is denoted as y_{ijk}, representing the covariance between assets i and j at time bin k. However, implementation can be challenging.

[23] We thank an anonymous referee for raising this issue.

Inherent discontinuity between existence/non-existence of liquidity. Quotes missing, or "liquidity evaporation" may occur perhaps due to certain fundamental reasons; there may exist something discontinuous in such a situation. It is a hard problem and beyond the scope of the current research to develop a theoretical microstructure model that produces theoretically sound and convincing explanations. The current statistical model does not take such an underlying structure into consideration; however, we could incorporate the feature by extending the model by utilizing *Torbit-type* modeling.

Dynamic modeling of market quality measures. The current approach does not explicitly model dynamics of intraday liquidity nor volatility. However, the time of the day information is treated in a static manner, specifically through the observed responses y_{ij} as well as the predictors x_{ij} and z_j. So, we could possibly draw some useful insights about the dynamic relationships among, say, liquidity, bid-ask spreads and price changes by carefully examining results obtained by the current model. We postpone the task to future work.

Other selections of variables. There is substantial room for variable selection regarding x_{ij}, w_i, z_j. For instance, widening of bid-ask spreads may reduce liquidity, which can concern some institutional investors. So it might be reasonable to introduce a dummy variable which takes the value of 1 when the spread exceeds a pre-specified threshold value. Besides, volatility on the specific return (individual asset return minus the market return) might also be a candidate, in place of the two volatilities included in the current model as distinct predictors. Variables about cancellation (such as order-to-cancellation ratio) can be useful as well. In a preliminary analysis, we included volatility of the target stock (based on trade prices) in x_{ij}; logarithmic market return, market volatility, market dollar volume in logarithms in z_j, where as proxies of the first two we used a logarithmic return and volatility of a liquid Nikkei 225-linked ETF (code 1321), while as a proxy of the last we used the total transaction volumes of the M stocks in logarithms. In any case we have to obey the following simple "principle:" If the objective of the analysis is to improve the goodness of fit in the training sample or to facilitate ease of interpretation, there are virtually no restrictions on candidacy. However, if prediction of y_{ij} is the objective, careful selections of the predictor variables have to be made regarding the timing of their availability.

Alternative measures of volatility. For computing realized volatility, we may use micro-prices in place of mid-quote prices in order to reflect activities involving the best quotes (e.g., Hayashi (2017)). Mid-quotes can be motionless ("stale price") not only for illiquid stocks but also for very liquid stocks i.e., if the stock has extremely thick orders placed on the best quotes. This phenomenon can often be observed for those with a large tick size compared to the price, e.g, Mizuho FG (code 8411) prior to July 22nd of 2014. Alternatively to realized volatility, we may adopt range-based volatility such as Parkinson (1980), Garman and Klass (1980); Rogers and Satchell (1991). These types use only the highest and lowest prices during the measurement period for volatility, possibly with open and close prices added to improve accuracy, whence they are immune to market microstructure noise contained in high-frequency data, due to price discreteness and bid-ask bounce. Besides, in order to compute them one needs to keep only a few or several values over the preceding period, hence it can save both memory storage and computational time. This salient feature yields a great advantage in the current context. We will consider their use in the future.

Limitation of the approach. As elaborated, our approach can handle datasets with various degrees of sparsity or scarcity, seamlessly from "warm-start" to "cold-start" situations. In particular, it can in principle

apply to datasets containing lots of illiquid stocks virtually with no modification. However, in the above analysis we only used stocks taken out of most liquid 500 stocks listed on the TSE. This is simply because, if we instead choose to use illiquid stocks for the analysis, then we could rarely observe the "true" values thus would not be able to evaluate the overall prediction accuracy of the model. This situation may contrast with usual applications of recommender systems such as movie reviews. There, the system can trace the target user *a posteriori* to check whether she/he has followed its recommendation such as viewing the recommended movie or purchasing the recommended item. In our context, however, *a posteriori* evaluation of the performance of the approach is essentially impossible, which is certainly a limitation and a challenge of the study.

4 Conclusion

This paper examines a methodology for market quality evaluation ("estimation") in the high-frequency domain – over short-time period – of individual stocks, by applying collaborative filtering.

Specifically, we adopted the regression-based latent factor model (RFLM) proposed by Agarwal and Chen (2009). This approach addresses the known weakness of the standard collaborative filtering methods, namely the "cold-start problem." It is a "hybrid"-type collaborative filtering method that can be used even when the observation data is very sparse, by performing regression using exogenous variables representing the characteristics of individual stocks and time zones as explanatory variables.

As the concrete market quality measures, we focused on liquidity and volatility. For liquidity, the ILOBS proposed by Deuskar and Johnson (2011) was adopted as the liquidity measure to be evaluated in this study. In the meantime, the realized volatility based on tick-by-tick mid-quote returns was adopted for the volatility measure. These choices were made mostly for an illustrative purpose, and irrelevant to the framework under investigation, hence other choices for market quality measures are possible.

In order to confirm the effectiveness of the methodology, empirical analysis was conducted using high-frequency limit-order book data obtained from the Tokyo Stock Exchange. Ninety-eight stocks out of the TOPIX 100 ingredients and ninety out of the Mid400 ingredients were used in the analysis. The data period ranges from January 4, 2019 to March 29, 2019 (58 business days), with data aggregated over every 5-minute interval.

As a result of the analysis, it was found that various factors that characterize market quality can be identified by the investigated methodology. The prediction results show that our approach performs stably and reasonably well compared to alternative models regardless of the percentage ratio of missing values, whence suggesting its robustness to the degree of sparsity or scarcity of observation data as intended. In the future we need to extend the data period and coverage to draw general empirical findings, and also to improve the methodology for better prediction accuracy.

Acknowledgements

We are grateful to the editor and two anonymous referees for constructive comments and suggestions. We also thank Takayuki Morimoto for helpful discussion at Nippon Finance Association 28th Annual Conference, Takatoshi Ito, Masahiro Yamada, Wataru Ohta, and Kenichi Ueda for constructive comments at Summer Workshop on Economic Theory 2020, and participants in those conferences. This research has been partially supported by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (C) (Project No.JP16K03601), MEXT Grant-in-Aid for Scientific Research (A) (Project No. 17H01100), JST CREST Grant Number JPMJCR14D7, JSPS Grant-in-Aid for Scientific Research (B) (Project No. 16H036050), and Hosei University's Innovation Management Research Center

research project, "Statistical Analysis on Information Propagation in Financial Markets and Related Areas." Data has been provided by the Japan Exchange Group, Tokyo Stock Exchange.

Conflicts of interest statement

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Agarwal, Deepak K. and Bee-Chung Chen (2009) "Regression-based latent factor models," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*.

——— (2016) *Statistical Methods for Recommender Systems*: Cambridge Univ. Press.

Aggarwal, Charu C. (2016) *Recommender Systems*: Springer.

Amihud, Yakov (2002) "Illiquidity and stock returns: cross-section and time-series effects," *J. Financial Markets*, Vol. 5, No. 1, pp. 31 – 56.

Andersen, Torben G. and Tim Bollerslev (1997) "Intraday periodicity and volatility persistence in financial markets," *Journal of Empirical Finance*, Vol. 4, No. 2, pp. 115 – 158. High Frequency Data in Finance, Part 1.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and P. Labys (2001) "The Distribution of Realized Exchange Rate Volatility," *Journal of American Statistical Association*, Vol. 96, pp. 42–55.

Barndorff-Nielsen, Ole E. and Neil Shephard (2002) "Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models," *Journal of Royal Statistical Society, Series B*, Vol. 64, pp. 253–280.

Belomestny, Denis, Mathias Trabs, and Alexandre B. Tsybakov (2019) "Sparse covariance matrix estimation in high-dimensional deconvolution," *Bernoulli*, Vol. 25, No. 3, pp. 1901–1938, 08.

Berry, William D. and Stanley Feldman (1985) *Multiple Regression in Practice (Quantitative Applications in the Social Sciences)*, California: SAGE Publications.

Bouchaud, Jean-Philippe, J. Doyne Farmer, and Fabrizio Lillo (2009) "How Markets Slowly Digest Changes in Supply and Demand," in Thorsten Hens and Klaus Reiner Schenk-Hoppé eds. *Handbook of Financial Markets: Dynamics and Evolution*, San Diego: North-Holland, Chap. 2, pp. 57 – 160.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2014) "High-frequency trading and price discovery," *Review of Financial Studies*, Vol. 27, No. 8, pp. 2265–2306.

Cont, Rama, Arseniy Kukanov, and Sasha Stoikov (2014) "The price impact of order book events," *J. Financial Econometrics*, Vol. 12, No. 1, pp. 47–88.

Dacorogna, Michel M., Ramazan Gençay, Ulrich A. Müller, Richard B. Olsen, and Olivier V. Pictet (2001) *An Introduction to High-Frequency Finance*, San Diego: Academic Press.

Deuskar, Prachi and Timothy C. Johnson (2011) "Market Liquidity and Flow-driven Risk," *Review of Financial Studies*, Vol. 24, No. 3, pp. 721–753.

Eisler, Zoltan, Jean-Philippe Bouchaud, and Julien Kockelkoren (2012) "The price impact of order book events: Market orders, limit orders and cancellations," *Quant. Finance*, Vol. 12, No. 9, pp. 1395–1419.

Fan, Jianqing and Donggyu Kim (2018) "Robust High-Dimensional Volatility Matrix Estimation for High-Frequency Factor Model," *Journal of the American Statistical Association*, Vol. 113, No. 523, pp. 1268–1283.

Fan, Jianqing, Yuan Liao, and Han Liu (2016) "An overview of the estimation of large covariance and precision matrices," *The Econometrics Journal*, Vol. 19, No. 1, pp. C1–C32.

Garman, Mark B. and Micahel J. Klass (1980) "On the Estimation of Security Price Volatility from Historical Data," *Journal of Business*, Vol. 53, No. 1, pp. 67–78.

Hasbrouck, Joel and Gideon Saar (2013) "Low-latency trading," *J. Financial Markets*, Vol. 16, No. 3, pp. 647–679.

Hautsch, Nikolaus and Ruihong Huang (2012) "The market impact of a limit order," *J. Economic Dynamics & Control*, Vol. 36, No. 4, pp. 501–522.

Hayashi, Takaki (2017) "Statistical analysis of high-frequency limit-order book data: On cross-market, single-asset lead-lag relationships in the Japanese stock market (in Japanese)," *Proceedings of the Institute of Statistical Mathematics*, Vol. 65, No. 1, pp. 113–139.

Kim, Donggyu, Yi Liu, and Yazhen Wang (2018) "Large volatility matrix estimation with factor-based diffusion model for high-frequency financial data," *Bernoulli*, Vol. 24, No. 4B, pp. 3657–3682, 11.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017) "The Flash Crash: High-Frequency Trading in an Electronic Market," *J. Finance*, Vol. 72, No. 3, pp. 967–998.

Kyle, Albert S. (1985) "Continuous Auctions and Insider Trading," *Econometrica*, Vol. 53, No. 6, pp. 1315–1335.

Lee, Charles M. C., Belinda Mucklow, and Mark J Ready (1993) "Spreads, depths, and the impact of earnings information: An intraday analysis," *Review of Financial Studies*, Vol. 6, No. 2, pp. 345–374.

Litzenberger, Robert, Jeff Castura, and Richard Gorelick (2012) "The Impacts of Automation and High Frequency Trading on Market Quality," *Annual Review of Financial Economics*, Vol. 4, No. 1, pp. 59–98.

McInish, Thomas H. and Robert A Wood (1992) "An Analysis of Intraday Patterns in Bid/Ask Spreads for NYSE Stocks," *The Journal of Finance*, Vol. 47, No. 2, p. 753–764, June.

Menkveld, Albert J. (2013) "High frequency trading and the new market makers," *J. Financial Markets*, Vol. 16, No. 4, pp. 712–740.

——— (2016) "The Economics of High-Frequency Trading: Taking Stock," *Annual Review of Financial Economics*, Vol. 8, No. 1, pp. 1–24.

Parkinson, Michael (1980) "The extreme value method for estimating the variance of the rate of return," *Journal of Business*, Vol. 53, No. 1, pp. 61–65.

Ricci, Francesco., Lior Rokach, and Bracha Shapira eds. (2015) *Recommender Systems Handbook*: Springer, 2nd edition.

Rogers, L.C.G. and Stephen E. Satchell (1991) "Estimating variance from high, low and closing prices," *Annals of Applied Probability*, Vol. 1, No. 4, pp. 504–512.

Takahashi, Makoto (2018) "J-GATE and Intraday Liquidity of Nikkei 225 Futures Market (in Japanse)," *Futures-Options Report*, Vol. 30, No. 3, pp. 1–5.

Wood, Robert A, Thomas H. McInish, and J. Keith Ord (1985) "An Investigation of Transactions Data for NYSE Stocks," *The Journal of Finance*, Vol. 40, No. 3, pp. 723–739, July.

Yano, Makoto (2009) "The Foundation of Market Quality Economics," *The Japanese Economic Review*, Vol. 60, No. 1, pp. 1–32.

法政大学イノベーション・マネジメント研究センター
The Research Institute for Innovation Management, HOSEI UNIVERSITY

〒102-8160 東京都千代田区富士見 2-17-1
TEL: 03(3264)9420　FAX: 03(3264)4690
URL: http://riim.ws.hosei.ac.jp
E-mail: cbir@adm.hosei.ac.jp