IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 14, 2017, accepted January 25, 2018, date of publication February 2, 2018, date of current version March 28, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2801263

Discovering the Trading Pattern of Financial Market Participants: Comparison of Two Co-Clustering Methods

GUANGWEI SHI, LIYING REN, ZHONGCHEN MIAO, JIAN GAO, YANZHE CHE, AND JIDONG LU
Shanghai Financial Futures Information Technology Co., Ltd, Shanghai 200122, China
Corresponding author: Jidong Lu (lujd@cffex.com.cn)

ABSTRACT Co-clustering is rapidly becoming a powerful data analysis technique in varied fields, such as gene expression analysis, data and web mining, and market baskets analysis. In this paper, two co-clustering methods based on smooth plaid model (SPM) and parallel factor decomposition with sparse latent factors (SLF-PARAFAC) are respectively applied to synthetic data set and investors' transaction-level data set from the China Financial Futures Exchange. We present the comparison between two methodologies. Both SLF-PARAFAC and SPM are efficient, robust, and well suited for discovering trading ecosystems in modern financial markets. We recognize temporal pattern differences of various trader types. The results help to develop a thorough understanding of trading behaviors, and to detect patterns and irregularities.

INDEX TERMS Co-clustering, trading behavior, time-series, machine learning.

I. INTRODUCTION

Nowadays, electronic markets have become popular venues for different types of financial assets, such as stocks, commodities, options, and futures, etc.. The introduction of electronic trading has reduced the floor's advantage by increasing anonymity and obfuscated the roles, relationships, and designations in the original trading ecosystems. The ever-increasing complexity of financial trading and markets, makes it difficult to consider the basic problems of recognizing and categorizing market participants, for example, fully understand the present effect of various participants on financial markets, let alone to develop policies and regulatory interventions that are robust to developments of financial ecosystem. Although there have been various attempts in the literature to classify traders' behaviors over a feature space containing the summary statistics and derived variables, it still has some issues in real life financial scenarios. For example, the sample is usually characterized by high-dimensional features; it is necessary to work with new feature selection approach [1], [2] in data mining to eliminate noisy and irrelevant features and enhance the feature quality.

In the financial market, usually only certain traders participate in the specific transaction pattern, and only some trading characteristics of certain traders in a time window will reflect the trading behavior patterns to be investigated. For example, high frequency trading accounts may engage in establishing and liquidating positions in very short timeframes when financial markets around the world plunge and fluctuate wildly. Therefore, to accurately characterize the roles and functions of the market participants, it is necessary to identify the trader groups that participate in certain pattern characteristics or the pattern characteristics associated with certain trader groups. Three-way clustering can be seen as a ''local" clustering, which can be determined by a subset of traders in a temporal pattern, or by a subset of the characteristics of certain trader group. Transaction-level data encompasses rich and extensive information on trading activities. Mining these data may reveal insights into trading mechanisms. [3] presents a method referred as the smooth plaid model to designate traders into five distinct categories which are consistent with the results recovered manually by Kirilenko *et al.* [4].

Another co-clustering method, SLF-PARAFAC, proposed by [5] is designed for multi-way data and implements a modified version of k-means based on multilinear decomposition with sparse latent factors. The imposition of latent sparsity allows stable identification of a great many overlapping co-clusters in the data.

In this study, to validate SLF-PARAFAC and compare it with the smooth plaid model, we employ two methods on both synthetic data and transaction data in china financial futures market and analyzes different co-clusters to prove the

meaningfulness of co-clusters in the financial market ecosystem. This study first explores the sequential trading patterns of different participants in the Chinese financial derivatives market.

The structure of this paper is organized as follow: We start with an overview of related works in Section II. In Section III, we briefly summarize two methodologies, SPM and SLF-PARAFAC. Section IV is devoted to applications and case studies. We conclude with a discussion of this study in the last section.

II. RELATED WORKS

Clustering analysis is an important tool for statistical analysis, which is widely recognized and well implemented in a variety of scientific fields, including pattern recognition, signal processing, and biological information, etc. [6]. Regular clustering algorithms discover groups based on the data of full features. This process comes with several limitations. First of all, when the clustering algorithm is used to express clustering data, to measure the similarity between the objects, we should use the expression level of the object under all the attributes, or feature clustering takes advantage of the level of expression of all objects under the attribute. A method that treats all objects or attributes equally will make it impossible to find all similar groups of objects in the expression matrix data [7], [8]. Secondly, because the traditional clustering algorithm can only classify the objects into a cluster, and thus artificially conclude that each object has only a single function, this conflicts with the actual situation. In reality, a large number of objects assume different functions under different characteristics or different time level, which means they have different path patterns. In this area, the most representative applications include identification of distinctive checkerboard patterns in gene expression microarray [9] and information retrieval and text mining of document subgroups with similar properties relative to subgroups of attributes [10]. A series of efficient formulations for sub-matrix analysis have been proposed in [9] and [11]–[14].

Classical clustering is in a single dimension, called one-way clustering, clustering only rows (objects) or columns (attributes), and cannot cluster rows and columns simultaneously. Two-way clustering (bi-clustering), also known as subspace clustering or co-clustering, is clustering the rows and columns of the matrix at the same time. That is to say, co-clustering methods will cluster objects and properties simultaneously. The concept of two-way clustering was proposed by Hartigan [15] in 1972, and in 2000 by Cheng and Church introduced into the analysis of gene expression profiles. This technique has played a key role in the gene expression analysis, such as clustering the two modes of the gene-condition matrix simultaneously, and extracting a subset pair as a co-cluster consisting of a subset of genes and a subset of conditions with significant correlations. Thus such a co-cluster unravels a particular pattern.

Nowadays, multi-way arrays, such as color images ([row, column, color]) [16], [17], microarray data ([gene, condition, time]) [18], [19], pipe failure records ([pipe, attributes, time]) [20], scientific impact ([publication, citation, time]) [21], [22] are widely emerged in many tasks of real-world clustering analysis, demanding effective techniques that can facilitate the detection of meaningful co-clusters hidden in such data sets [5], [23]–[25]. [24] proposed a method based on a relation graph model. Another method developed in [5] to detect possibly overlapping co-clusters is based on multi-linear decomposition with sparse latent factors.

III. METHODS

An unsupervised multidimensional factor analysis technique called the SLF-PARAFAC model is implemented in this study. If $\underline{X} \in \mathbb{R}^{I\times J\times N}$ is a given three-way tensor, its PARAFAC decomposition [26] in K rank-one components is defined as the sum of outer products of three vectors:

$$\underline{X} \cong \sum_{k=1}^{K} a_k \circ b_k \circ c_k \tag{1}$$

where $a_k \in \mathbb{R}^{I\times 1}$, $b_k \in \mathbb{R}^{J\times 1}$, $c_k \in \mathbb{R}^{N\times 1}$ and $a_k \circ b_k \circ c_k(i,j,n) = a_k(i)b_k(j)c_k(n)$. Alternately, the PARAFAC decomposition can be represented by the factor matrices $A \in \mathbb{R}^{I\times K}$, $B \in \mathbb{R}^{J\times K}$, $C \in \mathbb{R}^{N\times K}$ containing vectors a_k, b_k and c_k as columns respectively.

Paper [5] shows that sparsity on the latent factors of PARAFAC can help multi-way co-clustering to eliminate noise and separate overlapping co-clusters correctly. This can be implemented by adding penalties on the l_1 norm of the factors:

$$\begin{aligned}\min_{\{0\le\rho_k\le\bar{\rho},0\le a_k,b_k,c_k\le 1,k=1,\cdots,K\}} &\| \underline{X} \\ &- \sum_{k=1}^{K} \rho_k a_k \circ b_k \circ c_k \|_F^2 + \lambda_a \sum_k \| a_k \|_1 \\ &+ \lambda_b \sum_k \| b_k \|_1 + \lambda_c \sum_k \| c_k \|_1 \end{aligned} \tag{2}$$

where $\bar{\rho} = \max_{i,j,n} X(i,j,n)$ and $\|\cdot\|_F^2$ is the squared Frobenius norm. The minimum is taken over only those (a_k, b_k, c_k) that are either both zero or both nonzero. It is explained in [5] that having only one zero vector cannot be optimal. Notice that any $a_k \ge 0$ can be written as $\sigma_k \tilde{a}_k$, with $0 \le \tilde{a}_k \le 1$ and $\sigma_k := \max_i a_k(i)$; and likewise for b_k and c_k. Hence it makes sense to penalize $\| \tilde{a}_k \|_1$ instead of $\| a_k \|_1$ while imposing sparsity on $a_k > 0$ and retaining scaling freedom simultaneously.

If X and the latent factors have real-valued elements, then the problem can be formulated as

$$\begin{aligned}\min_{\{|\rho_k|\le\bar{\rho},-1\le a_k,b_k,c_k\le 1,k=1,\cdots,K\}} &\| \underline{X} \\ &- \sum_{k=1}^{K} \rho_k a_k \circ b_k \circ c_k \|_F^2 + \lambda_a \sum_k \| a_k \|_1 \\ &+ \lambda_b \sum_k \| b_k \|_1 + \lambda_c \sum_k \| c_k \|_1 \end{aligned} \tag{3}$$

TABLE 1. Summary of bi-cluster structure for the illustrative example.

Bi-cluster	$\mu_k^{(t)}$	Size	Rows	Columns
1	$\sqrt{t}$	11×11	20–30	20–30
2	2+cos(t)	31×31	10–40	10–40
3	-t/4	6×18	55–60	68–85
4	-2$\mathbb{1}${t>5}	26×15	5–30	53–67

where $\breve{\rho} = \max_{i,j,n} |X(i,j,n)|$ and vectors a_k, b_k and c_k are normalized to unit norm. Furthermore, [5] shows that sparsity produces additivity property of co-clusters.

Another type of three-way co-clustering analysis to be implemented in this study is the smooth plaid model (SPM) of [3], which extends [27] and seeks up- or down-regulated time responses. If the data array $X^{(t)}$ at time t is represented as

$$X_{ij}^{(t)} = \mu_0^{(t)} + \sum_{k=1}^{K} \theta_{ijk}^{(t)} r_{ik}^{(t)} c_{jk}^{(t)}, \tag{4}$$

where $(\{r_{ik}^{(t)}, c_{jk}^{(t)}, \theta_{ijk}^{(t)}\})$ indicate whether a co-cluster is active at time t. The objective function is given by

$$\begin{aligned} \min_{\theta^{(T-W)},\cdots,\theta^{(T)}} & \sum_{t=T-W}^{T} \sum_{i=1}^{n} \sum_{j=1}^{p} (\hat{Z}_{ij}^{(t)} - \theta_{ijK}^{(t)} \hat{r}_{iK}^{(t)} \hat{c}_{jK}^{(t)})^2 \\ & + \lambda \sum_{t=T-W+1}^{T} \sum_{i=1}^{n} \sum_{j=1}^{p} (\theta_{ijK}^{(t)} \hat{r}_{iK}^{(t)} \hat{c}_{jK}^{(t)} \\ & - \theta_{ijK}^{(t-1)} \hat{r}_{iK}^{(t-1)} \hat{c}_{jK}^{(t-1)})^2, \end{aligned} \tag{5}$$

where W is the length of lookback time window to be considered and $\hat{Z}^{(t)}$ is the residual data array in a given time t.

SPM seeks co-clusters $\langle S, V \rangle$ where S and V denote a set of samples and a set of variables respectively such that all samples in S manifest a similar time response across all subjects in V. Such a method helps alleviate the effects of transient patterns and strengthen the monitoring of structural regularities in the data (see [3]).

In our study, we benefit greatly from the work of researchers who developed algorithms for two prescribed methods. Note that associated codes are publicly available.

IV. EXPERIMENTAL RESULTS

In this section, we first compare the SLF-PARAFAC approach with smooth plaid model corresponding to the synthetic data. We also assess the impact of various parameter settings of SLF-PARAFAC in the process. Next we turn to a transaction-level dataset, and finally, illustrate and validate the performance of two methods.

A. SYNTHETIC DATA

Two methodologies are firstly demonstrated with simulated data. Matrix observations are of size 100 × 100 with a background layer of constant mean $\mu_0^{(t)}$=10. Four co-clusters are implanted with structure summarized in Table 1. The samples are then contaminated by adding i.i.d. standard normal

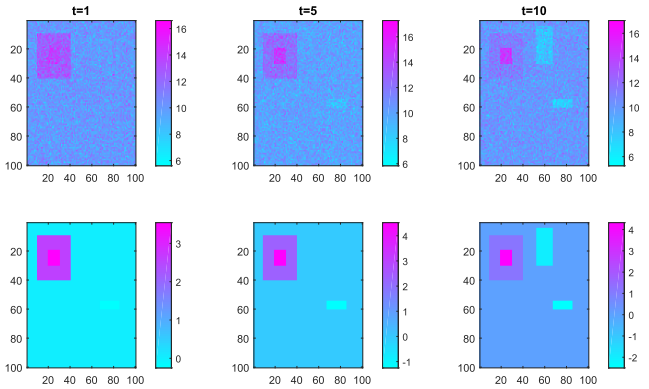


FIGURE 1. The upper panel shows samples of raw synthetic data. The lower panel shows samples of the filtered data (exclude the mean effects of background and noise effects).

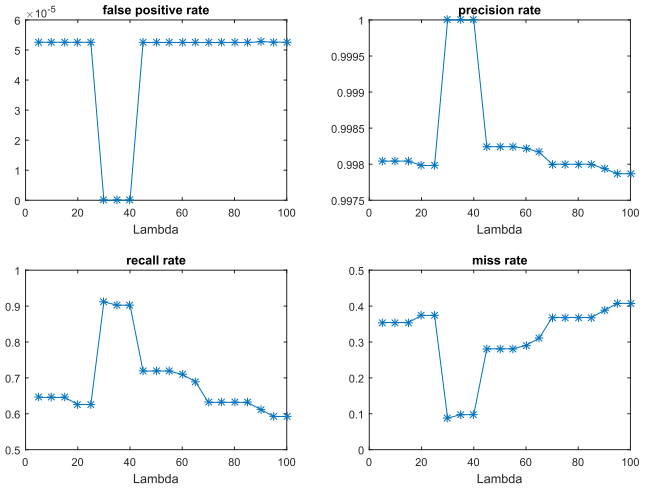


FIGURE 2. Different performance measures vs. λ with SLF-PARAFAC of synthetic data.

variables. Finally, sample the time uniformly between 1 and 10 and achieve 10 observed slices ($t = 1, 2, \cdots, 10$).

Table 1 illustrates examples of the raw data. As obviously seen, these sets overlap highly in both space and time. Table 2 shows the estimated co-clusters effects. We report results from running the smooth plaid model for three times. These results are labeled as SPM-1, SPM-2, and SPM-3 respectively. We also present results from setting the parameters for SLF-PARAFAC to equal 30. We see that two methodologies can reveal the correct matrix groups.

Results of the experiments concerning the setting parameter are shown in Figure 2. We can see a clear boost in performance with between 30 and 40. SLF-PARAFAC obtains a performance above 90% over precision rate and recall rate.

B. TRANSACTION-LEVEL DATA

The China Financial Futures Exchange (CFFEX) launched the country's first stock index futures, the Chinese Stock Index (CSI) 300 index futures, on April 16, 2010. Only after five years, the China Securities Regulatory

TABLE 2. Comparison of classification results based on synthetic dataset.

Algorithm	Dynamic Bi-cluster Detected	% Precision Rate	% False Positive	% Recall Rate	% Miss Rate
SPM-1	4	100%	0%	100%	0%
SPM-2	4	100%	0%	94.76%	5.24%
SPM-3	4	88.15%	0.55%	75.32%	24.68%
SLF-PAR-1	4	100%	0%	91.20%	8.80%

Commission announced the volume of CSI 300 stock index futures reached 1.9 trillion on April 2015, and exceeded E-Mini S&P 500 index futures (average volume 47 billion), and became the world's largest stock index futures product.

We collect and extract the real dataset from the China financial futures market intraday, tick by tick transaction level data, containing the following variables: number of trades, trading volume, change in inventory, cumulative net inventory, inter-trade duration for each trader. We calculate all the variables within a fixed time period of 15 minutes. All variables are same as in [3] excepted for the defined time period as 600 transactions. We give a concise explanation of each variable as follows.

- Number of trades is recorded for each trader as the total number of transactions made in a given time period. Number of trades contains significant information on transaction management strategies of a trader, such as the execution time horizons.
- Trading volume for each trader is the sum of the total number of contracts transacted during a given time period. This variable measures the overall trading activity and contains information that is useful to infer trading motivations of a trader.
- Change in inventory is the difference between total open long (buy) position and short (sell) position in contracts held by a particular trader during a given time period. Change in inventory might be useful to show the risk exposure level of a trader accumulated during a time period.
- Cumulative net inventory held by a trader is computed by accumulating the net inventory from the market opening time to the end of the current time period. This variable measures the risk exposure level of a trader accumulated from the market opening time.
- Inter-trade duration is measured by the time (in seconds) between two consecutive transactions involving a given trader. Take inter-trade duration as the median if a trader has at least one transaction during a sample period, otherwise take 900 seconds.

1) CLASSIFICATION OF TRADERS

We use a summary version of the data expressed in a three-way array of size $5000 \times 5 \times 328$, comprising the transaction-level time series data over one month (spanning from March 2, 2015 to March 30, 2015) indexed by sample, variable, and time period. We first compress the data range using a logarithmic transformation: X is mapped to X' as follows:

$$X' = \begin{cases} \log(X) + 1 & \text{if } X > 0 \\ 0 & \text{if } X = 0 \\ -\log(-X) - 1 & \text{if } X < 0 \end{cases} \quad (6)$$

We then fit a SLF-PARAFAC model for $\lambda_a = \lambda_b = \lambda_c = 10$ and SPM with the bandwidth of 1 to extract the dominant co-clusters. Extracted co-clusters unravel different groups and their temporal profiles.

TABLE 3. Grouping traders with the China financial future market dataset based on SLF-PARAFAC.

SLF-PAR	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5
Number	4287	58	55	57	108
Trader type	Noisy	HFT	Uncertain	MFT	Uncertain

TABLE 4. Grouping traders with the China financial future market dataset based on SPM.

SPM	Clust. 1	Clust. 2	Clust. 3	Clust. 4	Clust. 5
Number	4657	52	73	187	173
Trader type	Noisy	HFT	HFT	MFT	HFT

Table 3 and 4 presents a summary of classification results for two methods.[1] Next we will use "Method-Ck" to represents the cluster k identified by SLF-PARAFAC or SPM methods, where $k = 1, \cdots, 5$.

As shown in Figure 3, we find SLF-PAR-C1 and SPM-C1 are in line with the characteristics of small/residual/noisy traders that trade at the lowest frequencies with the longest average inter-trade duration. On the other hand, SLF-PAR-C2 and SPM-C2,C3,C5 reflect the frequent trade character of HFTs and intra-day traders, and are therefore grouped into the HFT-like category. It can also be clearly seen from traders in SLF-PAR-C4 and SPM-C4 that the persistence is medium for trade intensity and classified as the medium-frequency-like traders (MFTs) group. Although SLF-PAR-C3, C5 lack any persistent pattern, their trading activity bifurcates between medium-frequency-like traders group and small-like traders group.

2) COMPARISONS OF METRICS OF TRADING ACTIVITY

This subsection provides a comparison of time-series trading activity among different trader types.

Our first dimension of trader behavior is defined as a trader's portion of the total market volume. This measure

[1]The views expressed in this paper are the authors' own and do not necessarily reflect the CFFEX's official policy.

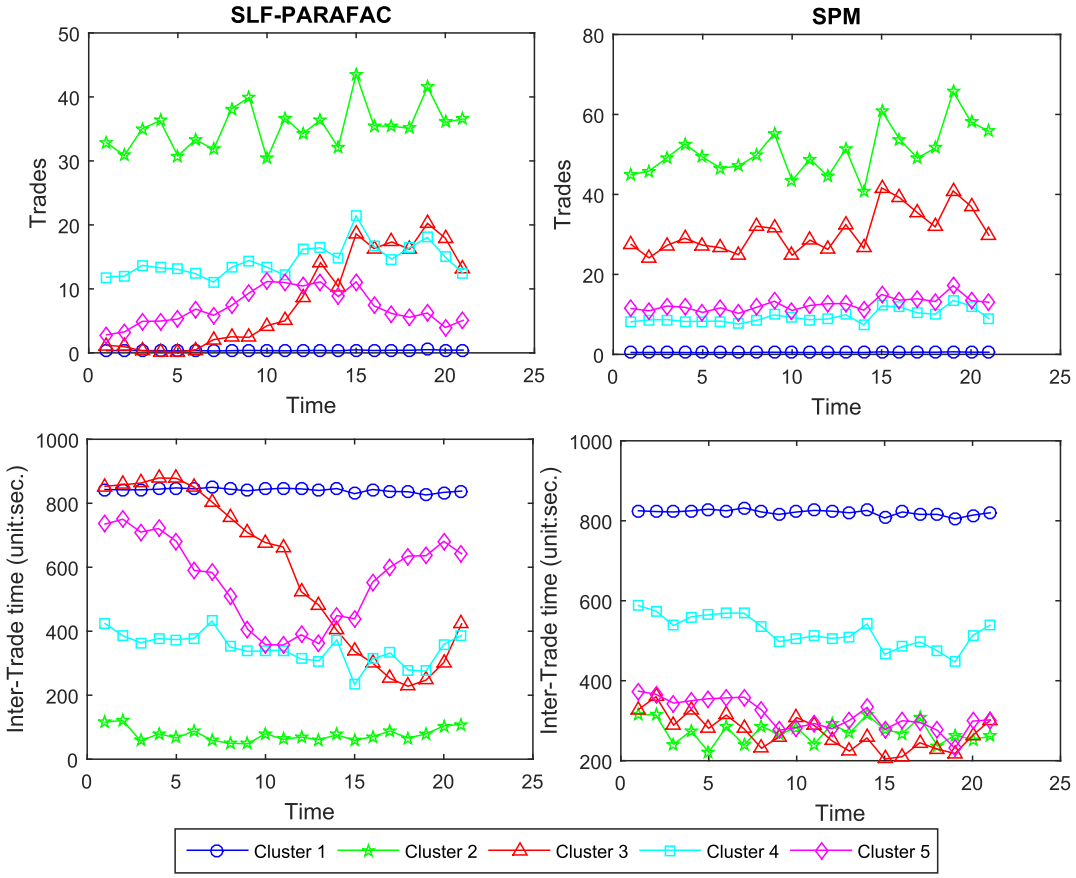


FIGURE 3. The upper panel portrays average trades for five trader groups over each day of the month with SLF-PARAFAC and SPM. The lower panel shows the average inter-trade durations with two models. Refer to Table 3 and 4 that each cluster represents which trader type.

is closely related to key features of trader activity. In general, high frequency traders (HFTs) alike are expected to be more active than other traders. [28] shows that the participation rate of HFT firms is 68.3% of dollar volume across the full sample of NASDAQ Datasets. In the E-Mini Datasets, HFT group containing 65 trading accounts produced 54.4% of market volume in [29] as compared to the 34.2% estimated in [4]. [30] indicates that 46.7% of volume during the period between September 17, 2010 and November 1, 2010 are generated by the designated HFT group of 30 accounts.

The second metric of trading activity is the absolute day-end inventory (the sum of all signed trading volumes) divided by daily trading volume, denoted by inventory/trade ratio. HFTs alike strive to have low inventory/trade ratio as positions held overnight are subject to clearing and capital costs.

Figure 4 shows that the market shares of various trader types range between 7.4% and 38.5%. Although there is no consensus on the precise share of HFT-like in overall trading volumes, it is very considerable for both two methods(ranking 1st in SLF-PARAFAC with 38.5% and the top three in SPM with 25.8%, 20.6% and 19.1%). Figure 4 shows HFT-like category has persistent lowest position compared to the trade volume. The average inventory/ trade among medium-frequency traders group ranks 1st and 2nd in SLF-PAR and SPM respectively. This group likely fits the fundamental category, and captures institutional investors, who are generally considered informed(see [31] and [32]).

In the study of U.S. financial derivatives market [3], traders can be designated into five persistent categories. The boundaries between the different types of participants are quite clear, and the statistical characteristics among the various types of traders are obviously at different levels.

3) IMPACT OF CHOICE OF SLF-PARAFAC MODEL PARAMETER λ

We show the impact of regularization parameter λ. For this purpose, we choose the transaction-level data. Figure 5 shows the number of iterations as well as running time versus λ using the transaction data. With the exception of a few occasional problems with local minima, the general result is clearly consistent with [5]: the number of iterations and running time until convergence are both decreasing functions of λ. This results from that the higher λ's are, more elements are zero-outed.

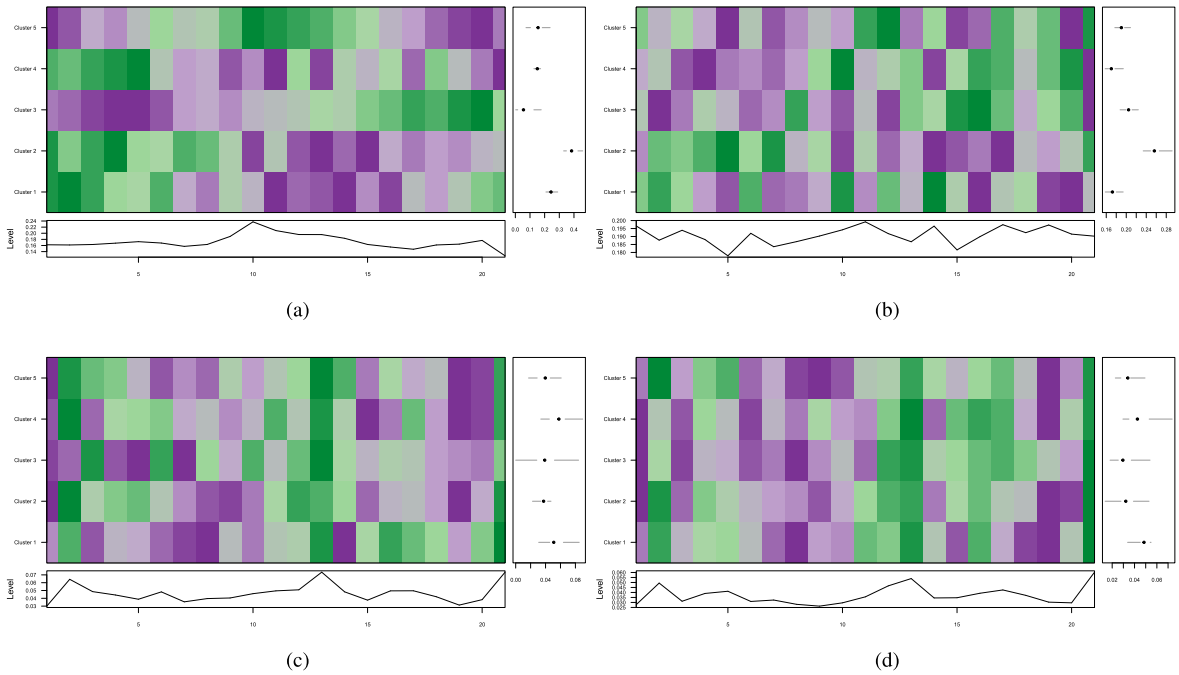


FIGURE 4. The upper panel shows average market volume shares for each group of traders over each day of the month with SLF-PARAFAC and SPM. The lower panel shows average inventory/trade ratio for each group of traders over each day of the month with SLF-PARAFAC and SPM. violet represents small values, and green represents big ones. (a) SLF-PARAFAC: market volume share. (b) SPM: market volume share. (c) SLF-PARAFAC: i/t ratio. (d) SPM: i/t ratio.

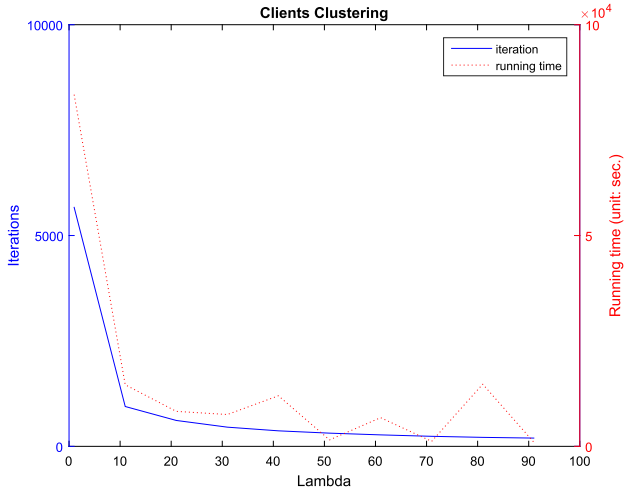


FIGURE 5. Number of iterations/running time vs. λ.

V. CONCLUSION

In this study, we present a brief description and comparison of co-clustering method and especially show how it can be applied to the classification of specific group of traders.

Both SLF-PARAFAC and SPM are effective methods on discovering overlapping co-clusters often occur in application. It is a challenging problem in clustering trader types in financial market. From the implementation of two methods on transaction-level data, it is seen that each trader may be designated into two or more categories which is difficult to recover by one-way clustering methods, such as k-means. These methods are scalable to large dataset, easy for application setting with a fewer tuning parameters.

SPM simply gives qualitative results, while SLF-PARAFAC can give quantitative results, indicating the degree of each elements belonging to a co-cluster. As a result, the application of SLF-PARAFAC is eventually a parameter design problem and it is necessary to determine threshold for the outputs to determine the membership between elements and categories.

From a regulatory point of view, a better picture of the actual scope of various market participants would be desirable. However, it is difficult for regulators to have comprehensive and accurate understanding of the ecosystem of financial market. The selection of indicators and standards is a complex multi-criteria problem including both quantitative and qualitative factors which may be uncertain or at most lack of robustness. This study supports the expression of various trader types in the financial market concerning trading practices in temporal pattern. We further conjecture that co-clustering is more informative and robust than summary statistic based approach, and it is appropriate for capturing new behavior patterns of market participants. SLF-PARAFAC and SPM, as two more advanced data-mining methodology recognizing persistent patterns in the multi-way data, are helpful for academic, policy and regulatory analysis.

Future work will mainly need to deal with bigger data sets generated by more features over different time periods and select relevant and important features for various market

participants. This will provide more better metrics to measure traders' behaviors and offer important insights to their latent effects on such important economic issues as the price formation processes and market quality, etc..

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous referees for their valuable comments. The views expressed in this paper are those of the authors and do not, in any way, reflect the views or opinions of the CFFEX, its Chairman, Managers or other staff. All the results are valid only for the sample data and not for the whole market, for reference only.

REFERENCES

[1] H. Yan, "Coclustering of multidimensional big data: A useful tool for genomic, financial, and other data analysis," *IEEE Syst., Man, Cybern. Mag.*, vol. 3, no. 2, pp. 23–30, Apr. 2017.

[2] C. Li, X. Wang, W. Dong, J. Yan, Q. Liu, and H. Zha. (2015). "Joint active learning and feature selection via CUR matrix decomposition." [Online]. Available: https://arxiv.org/abs/1503.01239

[3] S. Mankad, G. Michailidis, and A. Kirilenko, "Discovering the ecosystem of an electronic financial market with a dynamic machine-learning method," *Algorithmic Finance*, vol. 2, no. 2, pp. 151–165, 2013.

[4] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, "The flash crash: The impact of high frequency trading on an electronic market," Univ. Maryland, College Park, MD, USA, 2011.

[5] E. E. Papalexakis, N. D. Sidiropoulos, and R. Bro, "From K-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 493–506, Jan. 2013.

[6] R. Xu and D. Wunsch, II, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, May 2005.

[7] Z. Miao, J. Yan, K. Chen, X. Yang, H. Zha, and W. Zhang, "Joint prediction of rating and popularity for cold-start item by sentinel user selection," *IEEE Access*, vol. 4, pp. 8500–8513, 2016.

[8] G. Sherlock, "Analysis of large-scale gene expression data," *Current Opinion Immunol.*, vol. 12, no. 2, pp. 201–205, 2000.

[9] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.

[10] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1265–1287, Mar. 2003.

[11] B. S. Lam and H. Yan, "Subdimension-based similarity measure for dna microarray data clustering," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 4, p. 041906, 2006.

[12] S. Van Aelst, X. S. Wang, R. H. Zamar, and R. Zhu, "Linear grouping using orthogonal regression," *Comput. Stat. Data Anal.*, vol. 50, no. 5, pp. 1287–1312, 2006.

[13] X. Gan, A. W.-C. Liew, and H. Yan, "Discovering biclusters in gene expression data based on high-dimensional linear geometries," *BMC Bioinform.*, vol. 9, no. 1, p. 209, 2008.

[14] H.-C. Chen, W. Zou, Y.-J. Tien, and J. J. Chen, "Identification of bicluster regions in a binary matrix and its applications," *PLoS ONE*, vol. 8, no. 8, p. e71680, 2013.

[15] J. A. Hartigan, "Direct clustering of a data matrix," *J. Amer. Statist. Assoc.*, vol. 67, no. 337, pp. 123–129, 1972.

[16] P. Comon, "Tensors: A brief introduction," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 44–53, May 2014.

[17] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, 2009.

[18] L. Omberg, G. H. Golub, and O. Alter, "A tensor higher-order singular value decomposition for integrative analysis of dna microarray data from different studies," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 47, pp. 18371–18376, 2007.

[19] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, and O. Alter, "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms," *PLoS ONE*, vol. 6, no. 12, p. e28072, 2011.

[20] J. Yan *et al.*, "Towards effective prioritizing water pipe replacement and rehabilitation," in *Proc. IJCAI*, 2013, pp. 2931–2937.

[21] X. Liu, J. Yan, S. Xiao, X. Wang, H. Zha, and S. Chu, "On predictive patent valuation: Forecasting patent citations and their types," in *Proc. AAAI*, 2017, pp. 1438–1444.

[22] S. Xiao *et al.*, "On modeling and predicting individual paper citation count over time," in *Proc. 25th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2016, pp. 2676–2682.

[23] L. Zhao and M. J. Zaki, "TriCluster: An effective algorithm for mining coherent clusters in 3D microarray data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2005, pp. 694–705.

[24] A. Banerjee, S. Basu, and S. Merugu, "Multi-way clustering on relation graphs," in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 145–156.

[25] T. Wu, A. R. Benson, and D. F. Gleich, "General tensor spectral co-clustering for higher-order data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2559–2567.

[26] R. A. Harshman, "Foundations of the PARAFAC procedure: Models and conditions for an 'explanatory' multimodal factor analysis," in *Proc. UCLA Working Papers Phonetics*, vol. 16. 1970, pp. 1–84.

[27] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Stat. Sin.*, vol. 12, pp. 61–86, Jan. 2002.

[28] A. Carrion, "Very fast money: High-frequency trading on the NASDAQ," *J. Financial Markets*, vol. 16, no. 4, pp. 680–711, 2013.

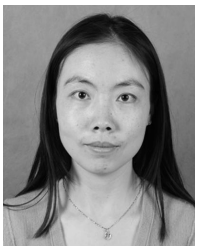
[29] M. D. Baron, J. Brogaard, and A. A. Kirilenko, "The trading profits of high frequency traders," *SSRN Electron. J.*, Jul. 2012. [Online]. Available: https://ssrn.com/abstract=2106158

[30] A. D. Clark-Joseph and A. Simsek, "Exploratory trading," Working paper, Univ. Illinois, Champaign, IL, USA, 2013.

[31] A. Boulatov, T. Hendershott, and D. Livdan, "Informed trading and portfolio returns," *Rev. Econ. Stud.*, vol. 80, no. 1, pp. 35–72, 2012.

[32] T. Hendershott, D. Livdan, and N. Schürhoff, "Are institutions informed about news?" *J. Financial Econ.*, vol. 117, no. 2, pp. 249–287, 2015.

GUANGWEI SHI received the B.S. degree in computer science from Chengdu Electronic Technology University, China, in 1993, and the M.S. degree in finance from the Shanghai University of Finance and Economics, China, in 2013. He is currently the Chief Executive Engineer with the Shanghai Financial Futures Information Technology Co., Ltd. His research interests include technical economics and management, credit risk management, and financial data mining.



LIYING REN received the B.S. and Ph.D. degrees in statistics and mathematical finance from Shandong University, Jinan, China, in 2005, and 2013, respectively. She joined the Shanghai University of Finance and Economics as an Assistant Professor in 2015. She is currently a Researcher with the Innovation Laboratory, Shanghai Financial Futures Information Technology Co., Ltd. Her research interests aim to use analytics and machine learning for financial modeling with time series data.

ZHONGCHEN MIAO received the B.S. and Ph.D. degrees from the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2010 and 2017, respectively. His research interests are information retrieval, user behavior analysis, machine learning, and data mining applications. He is currently the key member with the Institute on Investors' Behavior Patterns Research in financial markets.

JIAN GAO received the B.S. and M.S. degrees in computer software theory and application from Nanchang University, Nanchang, China, in 1999, and 2002, respectively. He manages the Machine Learning Systems Group of Shanghai Financial Futures Information Technology Co., Ltd. He has expertise in databases, data analytics, and distributed computing.

YANZHE CHE received the B.S. degree in software engineering and Ph.D. degree in computer science and technology from Zhejiang University, in 2006 and 2013, respectively. He is currently with the Shanghai Financial Futures Information Technology Co., Ltd. His research interests include machine learning and distributed databases technologies and applications.

JIDONG LU received the B.S. degree in communication engineering from Shanghai Tiedao University, Shanghai, China, in 1996. He is currently a General Manager with the Shanghai Financial Futures Information Technology Co., Ltd. His Department is currently involved in building systems for analytics, and large-scale machine learning. He has extensive experience consulting on financial data analytics and information services.