International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:15, No:4, 2021

Attribute Analysis of Quick Response Code Payment Users Using Discriminant Non-negative Matrix Factorization

Hironori Karachi, Haruka Yamashita

***Abstract*—** Recently, the system of quick response (QR) code is getting popular. Many companies introduce new QR code payment services and the services are competing with each other to increase the number of users. For increasing the number of users, we should grasp the difference of feature of the demographic information, usage information, and value of users between services. In this study, we conduct an analysis of real-world data provided by Nomura Research Institute including the demographic data of users and information of users' usages of two services; LINE Pay, and PayPay. For analyzing such data and interpret the feature of them, Nonnegative Matrix Factorization (NMF) is widely used; however, in case of the target data, there is a problem of the missing data. EM-algorithm NMF (EMNMF) to complete unknown values for understanding the feature of the given data presented by matrix shape. Moreover, for comparing the result of the NMF analysis of two matrices, there is Discriminant NMF (DNMF) shows the difference of users features between two matrices. In this study, we combine EMNMF and DNMF and also analyze the target data. As the interpretation, we show the difference of the features of users between LINE Pay and Paypay.

***Keywords*—**Data science, non-negative matrix factorization, missing data, quality of services.

I. Introduction

THE Japanese government has been lately promoting a cashless society, enabling people to purchase without paying cash. This has triggered rapid growth in the quick response (QR) code settlement market [1]. QR is a technology for representing information at high speed using the 2-dimensional binary codes. Recently, this technology realizes the payment vertically without the real money for purchases. QR code serves and store a large amount of information, not only numbers but also multilingual data. Various QR code payment services have been introduced, competing with each other. The market has now graduated beyond the introductory phase and entered a growth phase with companies aiming to further increase the user base by expanding the number of stores and launching reduction campaigns. However, the differences in attributes among the services and the features of users of each service have not yet been understood. Thus, examining the types of features among the users of each service is necessary. Therefore, in this study, we analyze the differences in the features of users who use these services.

Several feature analyses of data based on matrix factorization have been conducted [2]. In particular, the NMF method, which represents data as a product of two matrices when all elements have positive values and the features of data can be interpreted by focusing on the values of the matrix elements, has been widely used [3]. Applying the method to purchase history data enables us to understand the purchasing pattern of each user. Furthermore, many situations exist where the data include missing values. EMNMF, which is an NMF method in the case of missing values, has been proposed [4]. This method has been used by an e-commerce company to calculate the predicted evaluation values for unpurchased items by users utilizing the evaluation and purchase histories of similar users and to recommend items based on the ranking [4]. Besides, discriminative NMF (DNMF) has been proposed to clarify the features of each group, focusing on the case where data are divided into two groups [5]. This method has been applied to visualize the differences in the county information to improve the performance of image classification and face recognition [5]. Combining and analyzing these methods makes it possible to reveal the features of various types of data [6].

In this study, PayPay and LINE Pay, two QR code payment services, were chosen as the target services. PayPay is a service launched on October 5, 2018, by PayPay Corporation, a joint venture company between SoftBank Corp. and Yahoo Japan Corporation [7]. LINE Pay is a service launched in December 2014 by the LINE Corporation [8]. We analyze the questionnaire data provided by the Nomura Research Institute [9] for the two QR code payment services (PayPay and LINE Pay) by setting identical questions for each user and clarifying the differences between the two services to understand the features that exist among the users of each service. However, the data have many missing values, indicating the need to complement the values. Besides, it is desirable to analyze the two sets of data (i.e., PayPay and LINE Pay data), which consist of the same variables and represent the differences between them, using one model. In this study, we introduce the EMNMF framework [4] to complete the missing values for data analysis. Moreover, we use DNMF, which can clarify the difference between the two services using one model [5]. Furthermore, by visualizing the obtained results, we relate the features of the survey respondents to the features of the services. Besides, we propose marketing measures based on this analysis.

Hironori Kurachi is a student in the graduated school of science and engineering, Sophia University, 7-1 Kioicho, Chiyoda-ku Tokyo, 102-8554 (corresponding author, phone: +813-3238-3496; e-mail: hiro1216k1966@eagle.sophia.ac.jp).

Haruka Yamashita is an assistant professor in the school of science and engineering, Sophia University, 7-1 Kioicho,Chiyoda-ku Tokyo, 102-8554 (e-mail: h-yamashita-1g8@sophia.ac.jp).

II. EMNMF

EMNMF is a method of complementing missing values by approximating a matrix with missing values as a product of two low-dimensional nonnegative matrices [4]. Let the two matrices be the item set I = $\{I_i: 1 \leq i \leq N\}$, and the user set J = $\{J_j: 1 \leq j \leq M\}$. For item I_i, let A = $[a_{ij}] \in R^{N\times M}$ be a matrix whose components are the ratings y made by the user J_j and missing if not rated. Let W = $[w_{ik}] \in R^{N\times K}$ be the matrix indicating the affiliation of item I_i to the defined feature k$\{1 \leq k \leq K\}$, and let H =$[h_{kj}] \in R^{K\times M}$.Let X = $[x_{ij}] \in R^{N\times M}$be the approximation matrix of the item-user matrix calculated by the product of W and H (i.e., X = WH). A includes the set of elements for which values of A exist. We also consider the unobserved values. Let A^o be the set of elements for which a value of A exits and let A^u be the set of missing components. The approximation A' of A is learned by iterating (2)-(5). Let $A'^{(t)}, W^{(t)}, H^{(t)},$ and $X^{(t)}$ be the matrices after the t[th] update and $a'^{(t)}_{ij}, w^{(t)}_{ik}, h^{(t)}_{kj},$ and $x^{(t)}_{ij}$ be the (i,j) components of each matrix. The initial state is $t = 0$, and $\delta(\cdot)$ is an indicator function.

$$\min\|A' - WH\|_F^2 \text{ s.t. } \forall w_{ik}, \forall h_{kj} \geq 0 \quad (1)$$

$$w_{ik}^{(t)} = w_{ik}^{(t-1)} \frac{(A'^{(t-1)}H^{(t-1)T})_{ik}}{(W^{(t-1)}H^{(t-1)}T^{H^{(t-1)T}})_{ik}} \quad (2)$$

$$h_{kj}^{(t)} = h_{kj}^{(t-1)} \frac{(W^{(t-1)T}A'^{(t-1)T})_{kj}}{(W^{(t-1)}W^{(t-1)T}H^{(t-1)T})_{kj}} \quad (3)$$

$$a'^{(t)}_{ij} = a_{ij}\delta(a_{ij} \in A^u) + x_{ij}^{(t-1)}\delta(a_{ij} \notin A^u) \quad (4)$$

EMNMF has been incorporated in various studies, including matrix factorization-based enhancement filtering [10].

III. DNMF

DNMF incorporates the concept of linear discriminant into matrix factorization to identify the differences between groups [5]. Originally, DNMF was proposed to improve the discriminative ability of the matrix factorization in image classification. DNMF is a distance function based on Kullback-Leibler divergence [11], and a new term is added to it to incorporate the concept of linear discrimination. The distance function D_{DNMF} for DNMF is shown below.

$$D_{DNMF}(X, WH) = D(X, WH) + \alpha tr[S_w] - \beta tr[S_b] \quad (5)$$

To incorporate the concept of linear discrimination into the NMF distance function, we maximize Fisher's discriminant criterion. That is, the distance function in DNMF, D_{DNMF}, is a new term added to reduce the trace of the sum of the within-county covariance matrix S_W and increase the trace of the sum of the between-county covariance matrix S_b. S_w and S_b are obtained from the elements of the weight matrix H, which is estimated using matrix factorization.

$$S_w = \sum_{r=1}^{R}\sum_{p=1}^{Nr}(\boldsymbol{\eta}_p^{(r)} - \boldsymbol{\mu}^{(r)})(\boldsymbol{\eta}_p^{(r)} - \boldsymbol{\mu}^{(r)})^T \quad (6)$$

$$S_b = \sum_{r=1}^{R} Nr\left(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu}\right)\left(\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu}\right)^T \quad (7)$$

where $\boldsymbol{\eta}_p^{(r)}$ is defined as the p[th] column vector of the group r $(r = 1, \ldots, R)$ in the weight matrix, $\boldsymbol{\mu}^{(r)}$ is the average vector of $\boldsymbol{\eta}_p^{(r)}$, μ is the average vector of the whole county, and Nr is the number of columns of county r. Let α and β be constants that are set in advance by the analyst. Based on the above settings, the feature matrix W and weight matrix H in DNMF are obtained by solving the minimization problem.

$$\underset{s.t.\, w_{nk} \geq 0, h_{kj} \geq 0, \sum w_{nk} = 1(\forall k)}{minimize} D_{DNMF}(X, WH) \quad (8)$$

The equations to obtain the feature matrix W and weight matrix H are given below.

$$w_{ik} \leftarrow w_{ik} \frac{\sum_j x_{ij}h_{kj}}{\sum_j \sum_k w_{ik}h_{kj}k_{kj}} \quad (9)$$

$$h_{kj} \leftarrow \frac{T + \sqrt{T^2 + Uh_{kj}\sum_n w_{ik}\frac{x_{ij}}{\sum_{k'} w_{ik'}h_{k'j}}}}{\frac{1}{2}U} \quad (10)$$

The terms T and U are defined as follows.

$$T = (2\alpha + 2\beta)\left(\frac{1}{Nr}\sum_{\lambda,\lambda\neq m} h_{k\lambda}\right) - 2\beta\mu_k - 1 \quad (11)$$

$$U = 4\left(2\alpha - (2\alpha + 2\beta)\frac{1}{Nr}\right) \quad (12)$$

DNMF has been incorporated in various studies, for example, feature analysis in face recognition and image recognition [12].

IV. Data Analysis

A. Analyzed Data and Settings

Using the single-source data provided by the Nomura Research Institute [9], we analyze the "consumer purchasing behavior data." The number of variables used in this study was 23, comprising 11 variables related to personal attributes and 12 variables related to the channels used. The variables related to personal attributes were age, gender, marital status, number of children, household income, frequency of use of Social networking services (Twitter, Facebook, Instagram, and LINE), sensitivity to trends, and sociability. The variables related to the channels of use are the frequency of use of convenience stores (six types), supermarkets, department stores, drug stores, Internet shopping (smartphone), and fast food restaurants. To unify the scale of each variable, we convert the values of the variables to a range between 0 and 3.

For the sample used in this study, only the users who answered using the services (PayPay and LINE Pay) at least once a month and intended to use the services were selected.

We use DNMF to visualize the differences among service users and EMNMF to complete missing values and for comparison.

B. Determining the Number of Dimensions of the Feature Variables

To apply matrix factorization approaches, we set the number of dimensions of the feature variables. Therefore, we determine the number of dimensions of EMNMF and DNMF. First, let A be the matrix of the questionnaire data from 248 PayPay and 196 LINE Pay users, which are the samples with no missing values. Then, A′ is a matrix created by randomly selecting the same missing values as the data used (8%). For EMNMF, we plot the residual sum of squares of A′ and the matrix output by completing the missing values. The increase or decrease in the residual sum of squares when the number of dimensions is varied from 2 to 10 is depicted in Fig. 1.

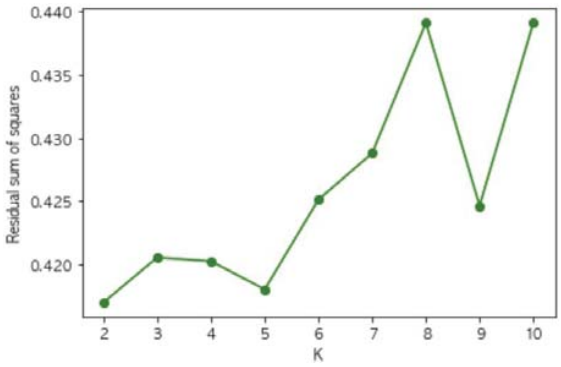

Fig. 1 Residual sum of squares for EMNMF varying in the number of dimensions (2 to 10)

From Fig. 1, we infer that the optimal number of dimensions is $K = 5$, where the residual sum of squares is small and I interpretability of results. Therefore, the dimension of EMNMF is set to five. By contrast, DNMF is applied to the two nonnegative matrices using A obtained by this method. The value of the residual sum of squares when the number of dimensions K is varied from 2 to 10 is illustrated in Fig. 2. As evident from Fig. 2, the optimal number of dimensions is $K = 3$, where the residual sum of squares is small.

C. Result of the Analyses Based on EMNMF and DNMF

We extracted only those samples for which the answer to the question on whether the respondent had used the services (PayPay and LINE Pay) at least once a month was in the affirmative and the response to whether they intended to use the services was either "want to use" or "definitely want to use." For the samples of 278 PayPay users and 229 LINE Pay users, we completed the missing values using EMNMF with the dimensionality set to five. Besides, the interpretation of the weight matrix W_1 and the feature matrix H_1 are obtained using DNMF with the dimensionality of three. The results are presented below.

We visualize the degree of cluster affiliation of each PayPay and LINE Pay user. To understand the features of each cluster, the feature matrix W_1 is illustrated in Fig. 3. As the number of users is large and it is difficult to represent the results in a tabulated form, we use color-coding, as depicted in Fig. 3.

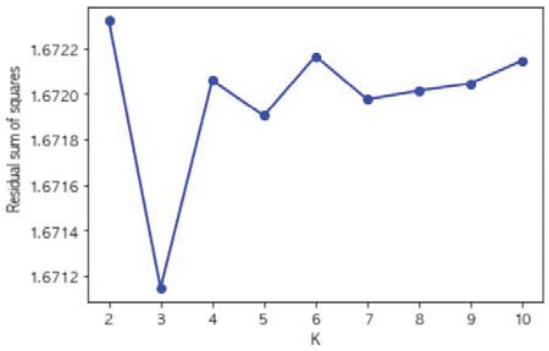

Fig. 2 Residual sum of squares for DNMF varying in the number of dimensions (2 to 10)

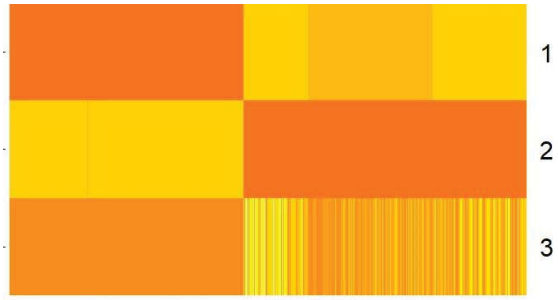

Fig. 3 Weight matrix W_1 by DNMF

The left half of the weighted matrix is for PayPay users, and the right half is for LINE Pay users. The darker the color, the higher is the degree of affiliation.

TABLE I
FEATURE MATRIX H_1 BY DNMF

	V1	V2	V3	V4	V5
Cluster1	0.047	0.048	0.065	0.054	0.035
Cluster2	0.056	0.044	0.063	0.054	0.035
Cluster3	0.005	0.039	0.059	0.001	0.036
	V6	V7	V8	V9	V10
Cluster1	0.044	0.024	0.034	0.041	0.060
Cluster2	0.044	0.025	0.032	0.040	0.062
Cluster3	0.135	0.005	0.019	0.045	0.022
	V11	V12	V13	V14	V15
Cluster1	0.052	0.050	0.046	0.050	0.058
Cluster2	0.053	0.050	0.048	0.046	0.057
Cluster3	0.003	0.031	0.026	0.181	0.004
	V16	V17	V18	V19	V20
Cluster1	0.060	0.036	0.036	0.034	0.037
Cluster2	0.056	0.033	0.039	0.038	0.035
Cluster3	0.020	0.023	0.031	0.134	0.033
	V21	V22	V23		
Cluster1	0.029	0.029	0.029		
Cluster2	0.030	0.030	0.030		
Cluster3	0.100	0.012	0.038		

Here, we introduce the elements of Table I as follows: convenience store A is Seven-Eleven, which is the leading

company in the convenience store industry. Convenience Store B is Lawson, which has three main types of stores: Regular Lawson, Natural Lawson, and Lawson Store 100, targeting different consumer segments; therefore, this brand has a rich variety of consumers. Convenience Store C is FamilyMart, which introduced the QR code payment service of PayPay that is more advanced than competing convenience stores. Convenience store D is Ministop, which is the only convenience store that offers fast food since the opening of its first store. Convenience Store E is Daily Yamazaki, which has a unique in-store cooking system that provides processed foods in the store.

We analyze the features of each service by calculating the degree of cluster affiliation of PayPay and LINE Pay users. We compare the difference between the two services by representing the ratio of the sample belonging to the weight matrix of PayPay and LINE Pay.

TABLE II
TABLE OF PAYPAY AFFILIATION RATE

	Cluster1	Cluster2	Cluster3
Affiliation rate	100%	0%	0%

TABLE III
TABLE OF LINE PAY AFFILIATION RATE

	Cluster1	Cluster2	Cluster3
Affiliation rate	0%	100%	0%

From the weight matrix, Clusters 1 and 3 are the feature matrices for PayPay users, whereas Cluster 2 is the feature matrix for LINE Pay users. We present below the interpretation obtained from the feature matrix H_1.

Features of PayPay

- Many male users;
- High number of older users;
- Many users at FamilyMart.

Features of LINE Pay

- Many female users;
- Many younger users;
- High Twitter usage rate;
- High usage rate of Seven-Eleven;
- High use of pharmacies and fast food.

The results show that the features of the matrix H_1 are obvious, and we can apply the results to marketing activities. We discuss the detailed marketing strategy in Subsection *E*.

D. Results of Analysis by Comparison Method

The EMNMF method was used for completing the missing values and for comparison. The interpretation of the resulting feature matrix H_2 is presented in Table IV. The number of dimensions of the feature variables was set to five.

We obtain the degree of cluster affiliation of PayPay and LINE Pay users.

Based on the weight matrix, the tendency of affiliation is similar for both PayPay and LINE Pay, and the difference between the services is not preferred. Focusing on the results of the comparison method, the analysis using DNMF is more effective in visualizing the differences between the services because it can understand the differences more clearly.

TABLE IV
FEATURE MATRIX H_2 BY EMNMF

	V1	V2	V3	V4	V5
Cluster1	2.692	2.047	0.000	2.494	0.000
Cluster2	1.050	1.517	5.794	1.904	1.948
Cluster3	0.000	1.806	0.000	1.751	0.000
	V6	V7	V8	V9	V10
Cluster1	1.668	1.403	1.655	1.970	2.909
Cluster2	0.000	0.449	0.000	0.232	0.082
Cluster3	3.109	0.000	2.050	2.645	4.421
	V11	V12	V13	V14	V15
Cluster1	2.932	2.482	2.401	0.000	0.581
Cluster2	0.000	0.194	0.076	0.461	0.000
Cluster3	3.447	3.421	3.233	4.471	5.875
	V16	V17	V18	V19	V20
Cluster1	0.000	0.001	2.946	2.472	2.846
Cluster2	0.453	0.000	0.837	0.699	0.605
Cluster3	6.072	2.193	0.157	0.328	0.247
	V21	V22	V23		
Cluster1	1.059	0.891	0.796		
Cluster2	0.222	0.121	0.056		
Cluster3	0.000	0.000	0.000		

TABLE V
TABLE OF PAYPAY AFFILIATION RATE

	Cluster1	Cluster2	Cluster3
Affiliation rate	27.34%	50.00%	22.66%

TABLE VI
TABLE OF LINE PAY AFFILIATION RATE

	Cluster1	Cluster2	Cluster3
Affiliation rate	34.50%	46.29%	19.21%

E. Consideration

Discussion on the Proposed Method and the Comparison Method

In the case of the proposed method using DNMF, the difference between the two services is represented by a single model, and the difference can be clarified with a small number of clusters. Therefore, the proposed method is effective for demonstrating the differences between multiple services because it can represent the differences between services more clearly.

Marketing Measures Based on the Results Obtained Using the Proposed Method

In this subsection, we present detailed marketing measures based on the analysis. First, as PayPay has many male users, we should improve the service to make it easier to use for this target group. Second, the high usage rate at FamilyMart can be due to the introduction of the PayPay QR code payment service that is more advanced than those in competing convenience stores. Therefore, it is suggested that the introduction of advanced PayPay leads to the increase in the use of convenience stores by QR code users. By contrast, there are three possible explanations for LINE Pay. First, as the ratio of young people to women is high, services should be improved so that they are suitable for young women. Second, because the

usage rate of Twitter is high, the use of Twitter-based advertisements is expected to be effective. Third, as the usage of pharmacies and fast food stores is high, the number of pharmacies and fast food stores as member stores introducing these systems should be increased.

V. Conclusion

In this study, we analyzed the differences in the features of the users of two QR code payment services (PayPay and LINE Pay) employing the two methods of matrix factorization, considering discrimination in the case of nonnegative matrices with missing values. We used EMNMF to supplement the missing values for data analyses, based on which, we used DNMF to visualize the differences among multiple services. Compared with the comparison method, we could visualize the differences clearly.

In contrast to the comparison method (i.e., EMNMF) we were able to further clarify the differences between the service users by using DNMF and demonstrated its effectiveness by analyzing the single-source questionnaire data provided by the Nomura Research Institute.

For future work, improving the model to classify equally for the set number of dimensions of each group is necessary because the information of clusters is biased.

Acknowledgment

In this study, we analyzed data provided by Nomura Research Institute Ltd. [9].

References

[1] Chen, Tietie; Yoko Ishino, 'Study on Popularization of QR Code Settlement in Japan.' Agents and Multi-Agent Systems: Technologies and Applications 2019. Springer, Singapore, 2020. 297-307.

[2] Berry, Michael W., et al. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 2007, 52.1: 155-173.

[3] Pauca, V. Paul, Jon Piper; Robert J. Plemmons. Nonnegative matrix factorization for spectral data analysis. *Linear Algebra & its Applications*, 2006, 416.1: 29-47.

[4] Zhang, Sheng, et al. Learning from incomplete ratings using non-negative matrix factorization. In: Proceedings of the 2006 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2006. 549-553.

[5] Zafeiriou, Stefanos, et al. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Transactions on Neural Networks*, 2006, 17.3: 683-695.

[6] Ano, Tsubasa, Haruka Yamashita; Masayuki Goto. An analysis of Consumer Panel Data Based on the Discriminant Nonnegative Matrix Factorization, Proceedings of the APIEMS2018, 2018. CD-included.

[7] Home page of PayPay Corporation, https://about.paypay.ne.jp/, last browsing: Feb 22, 2021.

[8] Com. Home page of LINE Corporation, https://LINEPaycorp/ja/, last browsing: Feb 22, 2021.

[9] Home page of Nomura Research Institute, Written in Japanese, https://www.nri.com/jp/, last browsing: 22nd Feb 2021.

[10] Ning, Shangbin; Fengchao Zuo. Sparsity-constrained NMF algorithm based on evolution strategy for hyperspectral unmixing. In: *MATEC Web of Conferences*. EDP Sciences, 2018, 232: 04019.

[11] Yang, Zhirong, et al. Kullback-Leibler divergence for nonnegative matrix factorization. In: *Lecture Notes in Computer Science* International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2011: 250-257.

[12] Nikitidis, Symeon, et al. Subclass discriminant nonnegative matrix factorization for facial image analysis. *Pattern Recognition*, 2012, 45.12: 4080-4091.