Received March 10, 2021, accepted March 16, 2021, date of publication March 19, 2021, date of current version April 2, 2021.
Digital Object Identifier 10.1109/ACCESS.2021.3067499

An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables

HONG-HAO ZHAO[1], XI-CHUN LUO[2], RUI MA[3], AND XI LU[1]
[1]Department of Decision Sciences, School of Business, Macau University of Science and Technology, Taipa 999078, China
[2]School of Intelligent Manufacturing, City College of Dongguan University of Technology, Dongguan 523000, China
[3]School of Computer Science, Zhuhai College of Science and Technology, Zhuhai 519000, China

Corresponding author: Hong-Hao Zhao (hhzhao@must.edu.mo)

This work was supported in part by the Faculty Research Grants of Macau University of Science and Technology under Grant FRG-19-040-MSB, and in part by the Department of Science and Technology of Guangdong Province under Grant 2020A0505090004.

ABSTRACT The omnichannel business has becomes a hot topic due to the fast development on e-commerce and the customers' acquaintance with multichannel shopping mode. Various business organizations have started to work on omnichannel business issue in order to satisfy the new trend of customer demand and tend to devote their efforts to both online and offline business. Thus, there is no doubt that understanding the shopping behavior for online customers is vital for the omnichannel business. The RFM (recency, frequency, monetary) model and the k-means clustering method are commonly used to extract customers' information and segment customers, respectively. To extend the RFM model, we divide the total frequency and monetary information into weekly level data, and as a consequence, the number of variables corresponding to one customer increases significantly, leading to the problem of high-dimensional analysis. To address this issue, in this paper we extend the regularized k-means clustering method with L_1-norm for independent case to the clustering method with elastic net penalty with a focus on correlated variables. Our simulation results show that the proposed method performs better than the standard k-means method by providing lower error rates and can select variables simultaneously under 4 different scenarios. A real example of an online retailer is presented to illustrate the use of the proposed method and highlight its high potential in clustering high-dimensional applications. In particular, the number of variables is reduced from 108 to 98 without any loss on clustering accuracy.

INDEX TERMS Customer segmentation, high-dimensional clustering, regularized K-means, correlated variables.

I. INTRODUCTION

The omnichannel business has begun to attract increasing attention from the public due to its wide range of applications and huge potentials in modern industry. It is well-known that traditional business activities focus on the offline transactions between business organizations and customers and the omnichannel integrates the traditional business with electronic business (E-business) and takes the advantages of both sides. With the rapid development of current internet technologies, E-business is experiencing a rapid growth [1]. There exists a clear trend that customers are switching their shopping mode from traditional outlets to internet, eliminating the time and geographical constraints on shopping [2], [3]. Furthermore, by shopping online, customers can acquire product information from various channels and tend to switch between different channels to obtain better shopping experience. To accommodate the new shopping pattern of the customers and provide better services, business organizations apply different strategies to achieve omnichannel business and one of the popular example is the online-to-offline (O2O) business mode.

It is clear that the successful implementation of omnichannel business relies on the correct understanding of customer online shopping preference. Customer segmentation is precisely the process of identifying different groups of customers according to their shopping behavior [4]–[6]. Unlike traditional customer segmentation that depends on variables such

The associate editor coordinating the review of this manuscript and approving it for publication was Zhan Bu.

as demographics and attitude [7], transaction records play an important role in online customer segmentation Fortunately, modern development in information technology allows the business organizations to track customers' transaction data precisely and thus segment the customers [8].

To apply customer segmentation, identification of the potential variables that could express the customers' shopping behavior becomes an essential task [9]. The RFM model proposed by Hughes [10] is a well-known model to characterize the customer shopping behaviors [11]. The RFM formulates the characteristics of customers by extracting the information such as the recency of the last purchase, frequency of the purchase and the monetary value of the purchase from the transaction records of the customers. The RFM model is used in many circumstances due to its simplicity, and successful applications can be found in Bult and Wansbeek [12], Newell [13], Migkautsch [14], Wu and Lin [15] and Lee and Park [16].

Based on the variables identified by the RFM model, data mining techniques can be applied to group the customers. Clustering is a type of data mining techniques that divides a set of objects into several categories where similar objects are grouped together([17], [18]). K-means is a well-known clustering algorithm that was first proposed by MacQueen [19]. It has been shown to be an effective clustering technique that can handle the clustering problem efficiently [20], although many other approaches have been reported in the literature [21]–[25]. Successful examples can also be found in Wei *et al.* [26], Mesforoush and Tarokh [20], Cheng and Chen [1] and Kuo *et al.* [27].

Although the RFM model and standard k-means method have obtained appealing results, the problems tend to be different under the modern business circumstance. Noted that the RFM model constructs three variables which are recency, frequency, monetary. Take shopping frequency as an example, the RFM model only calculate the total sum of the shopping frequency of each customer. Customers with the same total shopping frequency are not necessary to have similar shopping behavior, rather the shopping frequency distribution over time reflects the shopping pattern and behavior more precisely. Under such circumstance, to precisely segment the online customers, shopping frequency with a smaller time scale is required. Thus the traditional RFM model becomes ineffective due to the time scale it uses. To achieve a smaller time scale, one of the possible solution is to decompose the yearly data into weekly level data by extracting the customers' transaction data records, which is different from the traditional RFM model that applies the total sum to describe the shopping frequency and spending for each customer. Furthermore, other click-stream data available in the e-commerce environment could also be introduced. As a result, the number of variables which characterizes the customers' behavior increases significantly and leads to the problem of high-dimensional clustering.

Consequently, the standard k-means method becomes less effective in high-dimensional conditions [28]. The regularized k-means method that adds a penalty term to the standard k-means method can be an effective approach for handling the high-dimensional problem. Successful applications in the cluster high-dimensional case with independent variables are discussed by Witten and Tibshirani [29] and by Sun and Wang [30]. However, the variables cannot be purely independent in real applications and little research has been done along this direction.

In general, in the modern e-commerce environment, precise customer segmentation requires more variables which characterizes the customers' behavior. When small time scale data is decomposed from the total sum, it is possible that the number of customer features is more than the number of customers especially for small business and platform with very limited number of customers. As a result, the features used for customer segmentation are correlated and the problem becomes a high-dimensional clustering problem. Consequently, an effective clustering method for high-dimensional clustering with correlated variables is desired. Therefore, in this paper, we extend the regularized k-means clustering method from the focus on independent variables to the focus on correlated variables by introducing the elastic net penalty. The proposed method is assessed under high-dimensional clustering conditions with correlated variables. Our simulation results show that the proposed method performs better than the standard k-means method by providing lower error rates and can select variables simultaneously under 4 different scenarios. Furthermore, the number of variables is reduced from 108 to 98 without any loss on clustering accuracy in the application on real example.

The rest of the paper is organized as follows. In section 2, a detailed description of the proposed method is presented. In section 3, we conduct a simulation study to compare the performance of the proposed method and the standard k-means method. A real example application in online retail is provided to illustrate the use of our method in section 4. The concluding remarks and research limitations are mentioned in section 5.

II. DESCRIPTION OF METHODS

In this section, we give a brief review of the traditional k-means clustering method and introduce the proposed regularized k-means method. The proposed methods aim to provide sensible clustering results in high-dimensional applications with correlated variables. We note that such conditions can easily occur in the electronic business environment especially for small business or platform with limited number of customers.

A. K-MEANS CLUSTERING

Suppose we have a dataset X of n observations denoted as $X = (X_1, X_2, \ldots, X_n)$. And each $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ $(i = 1, \ldots, n)$ is a p-dimensional vector. The objective of k-means clustering is to segment the original data X into K prespecified clusters that minimize the total distance between the cluster center and the data within the corresponding cluster. The desired K clusters can be found by solving the following optimization problem,

$$\min_{B_j, C_j} \sum_{j=1}^{K} \sum_{X_i \in B_j} ||X_i - C_j||_{L_2}^2, \quad (1)$$

where $B_j (j = 1, 2, \ldots, K)$ and $C_j = (C_{j1}, C_{j2}, \ldots, C_{jp})^T (j = 1, 2, \ldots, K)$ are the K clusters and the corresponding cluster centers, respectively. $|| \cdot ||_{L_2}$ is the Euclidean norm or the L_2-norm. In this paper, we refer to this clustering method as the standard k-means method.

Finding the global optimal solution for Eq. (1) is an NP-hard problem. Therefore, Lloyd [31] proposed an iterative algorithm to approximate the solution for Eq. (1). The key idea of the algorithm is to update clusters B_j and centers C_j separately by assuming that the other variable is fixed during each iteration. In particular, the detailed algorithm is shown below.

Algorithm K-Means

Step 1. Initialize the centers $C_j^{(0)} (j = 1, 2, \ldots, K)$ by randomly choosing K observations from the original dataset X.
Step 2. Given centers $C_j^{(t-1)}$ at iteration $t-1$, find the clusters $B_j^{(t)}$ by assigning each observation X_i to the closest center.
Step 3. Given clusters $B_j^{(t)}$, update the centers $C_j^{(t)}$ by calculating the centers of observations $X_i \in B_j^{(t)}$. The detailed equation is given by $C_j^{(t)} = (\#B_j^{(t)})^{-1} \sum_{X_i \in B_j^{(t)}} X_i$, where $\#B_j^{(t)}$ is the cardinality of $B_j^{(t)}$.
Step 4. Repeat Step 2 and Step 3 until B_j is stable.

K-means clustering has been successfully applied in many fields. However, its clustering performance tends to be less effective in high-dimensional applications [28]. In addition, k-means clustering is not able to select the informative variables which is crucial in high-dimensional analysis.

B. REGULARIZED K-MEANS CLUSTERING

To address the drawbacks of the standard k-means method, we introduce the regularized k-means method in this section.

The idea of regularization is proposed by Tibshirani [32] where a penalty term is added to the original least squares function to obtain an estimate for the coefficients in the linear regression problem. The added penalized term can shrink the coefficients to zero when those coefficients are close to zero. As a result, the regularization method handles model fitting and variable selection simultaneously. Detailed properties of regularization method can be found in Zou and Hastie [33].

In this paper, we implement a similar approach by adding a regularization term to Eq. (1) similar to Sun and Wang [30]. The new optimization problem is expected to solve high-dimensional clustering problem by taking the advantage of the regularization term, and the detailed form is given by

$$\min_{B_j, C_j} \sum_{j=1}^{K} \sum_{X_i \in B_j} ||X_i - C_j||_{L_2}^2 + \sum_{m=1}^{p} P(C_{(m)}), \quad (2)$$

where $P(C_{(m)})$ is the regularization term added to each variable and $C_{(m)} = (C_{1m}, C_{2m}, \ldots, C_{Km})^T$ is the vector for the *mth* element for centers $C_j (j = 1, 2, \ldots, K)$.

We note that the regularization term $P(C_{(m)})$ can have various formats that provide the properties fitted for different applications. One of the commonly used penalty term is the L_1-norm penalty [34] that is given by

$$P(C_{(m)}) = \lambda_1 ||C_{(m)}||_{L_1}, \quad (3)$$

where $||C_{(m)}||_{L_1} = \sum_{j=1}^{K} |C_{jm}|$ and λ_1 is the tuning parameter that balances the sparsity and cluster model fitting. Other forms of using regularization terms can be found in Wang and Zhu [35] and Sun and Wang [30]. In this paper, we refer to this clustering method as the L_1 k-means method.

To better accommodate the correlation effects between different variables, in this paper we propose to use the elastic net penalty that combines the L_1-norm and L_2-norm penalty for which the detailed form is given by

$$P(C_{(m)}) = \lambda_2 \left\{ \frac{(1-\alpha)}{2} ||C_{(m)}||_{L_2} + \alpha ||C_{(m)}||_{L_1} \right\}, \quad (4)$$

where $||C_{(m)}||_{L_2} = \sum_{j=1}^{K} C_{jm}^2$ is the L_2-norm regularization term and α is the tuning parameter that balances the weights of the L_1-norm and the L_2-norm [33]. Finally, λ_2 is the other tuning parameter that has the same function as λ_1 in Eq. (3). The performance of elastic net penalty was assessed in Zou and Hastie [33] for regression purpose, but few studies have focused on its performance in clustering problems, particularly when correlations between the variables exist. Therefore, we propose to use the elastic new penalty in this paper, and we refer to this method as the L_{EN} k-means method. One should note that if λ_1 or $\lambda_2 = 0$, the regularization term $P(C_{(m)})$ becomes zero, and the regularized k-means methods are reduced to the classical k-means clustering method.

C. ALGORITHM FOR SOLVING REGULARIZED K-MEANS CLUSTERING

The clustering result of the regularized k-means method is the optimal solution for Eq. (2). Similar to the standard k-means, the optimal solution for the regularized k-means is not easy to achieve. Therefore, we still implement an iterative approach to solve Eq. (2). Similar to the iterative approach for solving the standard k-means method, we still calculate the clusters B_j and the centers C_j separately. When C_j is fixed, B_j can be easily generated by assigning X_i to the closest center. When the cluster B_j is fixed, C_j can be obtained in a componentwise fashion as discussed in Sun and Wang [30], and the transformation is given by

$$\min_{B_j, C_j} \frac{1}{n} \sum_{j=1}^{K} \sum_{X_i \in B_j} ||X_i - C_j||_{L_2}^2 + \sum_{m=1}^{p} P(C_{(m)})$$

$$= \sum_{m=1}^{p} \left\{ \frac{1}{n}(X_{(m)} - \Lambda C_{(m)})^T (X_{(m)} - \Lambda C_{(m)}) + P(C_{(m)}) \right\}, \quad (5)$$

where $X_{(m)} = (X_{1m}, X_{2m}, \ldots, X_{nm})^T$ is the vector that contains the mth variable element of all observations in X and Λ is an $n \times K$ cluster assigning matrix with elements in the form given by

$$\Lambda_{ij} = \begin{cases} 1 & X_i \in B_j \\ 0 & X_i \notin B_j. \end{cases} \quad (6)$$

As shown in Sun and Wang [30], when Λ in Eq. (5) is fixed, the optimal solution of Eq. (2) can be obtained by solving

$$\min_{C_{(m)}} \frac{1}{n}(X_{(m)} - \Lambda C_{(m)})^T (X_{(m)} - \Lambda C_{(m)}) + P(C_{(m)}) \quad (7)$$

for each $C_{(m)}$. Thus, the iterative approach for solving Eq. (2) is shown as follows.

Algorithm Regularized K-Means

Step 1. Initialize the centers $C_j^{(0)} (j = 1, 2, \ldots, K)$ by applying the standard k-means method to the observations from the original dataset X.
Step 2. Given centers $C_j^{(t-1)}$ at iteration $t-1$, find the clusters $B_j^{(t)}$ by assigning each observation X_i to the closest center and find the cluster assigning matrix Λ^t consequently.
Step 3. Given the cluster assigning matrix Λ^t, update the centers $C_j^{(t)}$ by solving Eq. (7) for each m.
Step 4. Repeat Steps 2 and 3 until Λ is stable.

We note that the cluster assigning result does not change if the cluster assigning matrix Λ^t is stable, and the iterative approach stops then. According to the simulation experiment and the results from Sun and Wang [30], the iterative approach stops within 5 iterations normally.

D. TUNING PARAMETERS SELECTION

As mentioned above, the proposed regularized k-means methods have several tuning parameters that must be specified prior to their implementation. In this section, we will discuss the major procedures for selecting the appropriate tuning parameters. We note that the L_{EN} k-means method has three parameters, namely, K, λ_2 and α, and the L_1 k-means method has two parameters, namely, K and λ_1. It is easy to observe that for any prespecified value of α, the parameter selection for L_1 k-means and L_2 k-means methods becomes exactly the same. Without loss of generality, we focus on the parameter selection of K, α and λ ($\lambda \in (\lambda_1, \lambda_2)$).

To measure the performance of different parameter settings, we must specify a criterion first. In this research, we follow the suggestions in Sun and Wang [30] and use the clustering stability as the evaluation criterion. The clustering stability can be described by the robustness for the clustering assignments given the same parameter settings. Such stability is obtained by measuring the dissimilarity or the distance between two clustering assignments. This means that a sensible choice of parameters is such as to guarantee that the clustering assignments calculated from different observations should have small distance in between, given that the observations are sampled from the same population.

Assume $Y = (X_1, X_2, \ldots, X_n)$ is the available dataset. Denote $\Gamma(Y|K, \alpha, \lambda)$ as the clustering assignment obtained through sample observation Y given parameter combination (K, α, λ). The parameter K, α and λ are chosen from the candidate parameter set given by $K \in (2, 3, \ldots, K_{max})$, $\alpha \in [0, 1]$ and $\lambda \in (\lambda \geq 0)$, respectively, where K_{max} is the maximum number of clusters considered and $K = 1$ is excluded since one cluster provides little information on customer behavior. To calculate the dissimilarity, we generate three bootstrap sample from the original dataset Y with the same sample size n, denoted as $\boldsymbol{Y_1^r}, \boldsymbol{Y_2^r}, \boldsymbol{Y_3^r}$ where $r = 1, 2, \ldots, R$ refers to the replications. Then, we generate the first two clustering assignments $\Gamma_1^r = \Gamma(\boldsymbol{Y_1^r}|K, \alpha, \lambda)$ and $\Gamma_2^r = \Gamma(\boldsymbol{Y_2^r}|K, \alpha, \lambda)$, and the dissimilarity for the rth replication is measured by the distance between Γ_1^r and Γ_2^r on sample $\boldsymbol{Y_3^r}$ that is given by

$$\begin{aligned} &D^r(\Gamma_1^r, \Gamma_2^r|\boldsymbol{Y_3^r}) \\ &\quad = (C_n^2)^{-1} \# \Big((i, j) \in S_n : \boldsymbol{I}\big(\Gamma_1^r(X_i^{(3)}) = \Gamma_1^r(X_j^{(3)})\big) \\ &\quad \neq \boldsymbol{I}\big(\Gamma_2^r(X_i^{(3)}) = \Gamma_2^r(X_j^{(3)})\big) \Big) \end{aligned} \quad (8)$$

where $\boldsymbol{I}(\cdot)$ is an indicator function, $C_n^2 = \frac{n(n-1)}{2}$, S_n is a set that contains all of the possible combinations of two observations from a population of size n and $\#(A)$ is still the cardinality of set A. Therefore, the optimal combination for K, α and λ is given by

$$(K, \alpha, \lambda) = \arg \min_{K, \alpha, \lambda} \sum_{r=1}^{R} D^r(\Gamma_1^r, \Gamma_2^r|\boldsymbol{Y_3^r}), \quad (9)$$

where D^r is a function of K, α and λ as already shown previously.

III. SIMULATION EXPERIMENT

In this section, we provide a simulation study to assess the capability of the proposed method and compare its performance with the standard k-means method in the high-dimensional case.

In the simulation, we provide a more general performance analysis compared with using real case example. The performance of each method is investigated under several scenarios, which avoids the bias from a specific dataset. Following similar logic in Sun and Wang [30], without loss of generality, we generate 20 observations with dimension p, denoted as $\boldsymbol{X} = (X_1, X_2, \ldots, X_{20})$ and their corresponding true cluster Z_i index is randomly generated from set $\{1, 2, 3, 4\}$. For each observation, the first 50 dimensions are the informative variables that are generated from $N(\mu(Z_i), \Sigma)$ where Σ is a 50 by 50 covariance matrix with element $\Sigma_{ij} = \rho^{|i-j|}$ and ρ is the factor that measures the correlation level. The mean

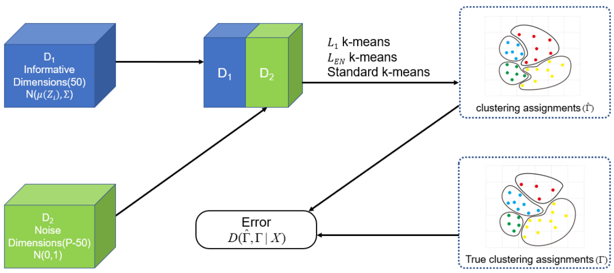

FIGURE 1. A brief summary for the simulation process.

vector $\mu(Z_i)$ is given by

$$\mu(Z_i) = \begin{cases} (-\mu\mathbf{1}_{25}^T, \mu\mathbf{1}_{25}^T)^T & (Z_i = 1) \\ (\mu\mathbf{1}_{50}^T)^T & (Z_i = 2) \\ (\mu\mathbf{1}_{25}^T, -\mu\mathbf{1}_{25}^T)^T & (Z_i = 3) \\ (-\mu\mathbf{1}_{50}^T)^T & (Z_i = 4) \end{cases} \tag{10}$$

where $\mathbf{1}_{50}$ is a vector of 50 ones. The remaining $p - 50$ variables are generated from $N(0, 1)$ that represent the noise term.

To assess the capabilities of different methods and compare their performance, the clustering error is measured by the distance between the estimated clustering assignment $\hat{\Gamma}$ and the true clustering assignment Γ on sample observation $\boldsymbol{X}$ that is given by

$$\begin{aligned} &D(\hat{\Gamma}, \Gamma|\boldsymbol{X}) \\ &= (C_n^2)^{-1}\#\Big((i,j) \in S_n : \boldsymbol{I}\big(\hat{\Gamma}(X_i) = \hat{\Gamma}(X_j)\big) \\ &\neq \boldsymbol{I}\big(\Gamma(X_i) = \Gamma(X_j)\big)\Big). \end{aligned} \tag{11}$$

And a brief summary for the simulation process is provided in Figure 1.

In our simulation, we investigate the performance of each method under different parameter values. In particular, we choose $p = 60, 80, 100$, $\mu = 0.4, 0.6, 0.8$ and $\rho = 0.6, 0.7, 0.8, 0.9$. Furthermore, since true cluster number K is crucial for the performance on each method, we examine each method in two scenarios. Scenario 1 assumes K is known, and scenario 2 assumes K is unknown.

A. CLUSTERING RESULTS WITH KNOWN K

In this simulation, the true number of cluster K is known, so that we fix $K = 4$ for all clustering methods. For the regularized k-means methods, the candidate set for tuning parameters λ_1 and λ_2 is $\{0.05\Delta; \Delta = 1, 2, \ldots, 10\}$. The optimal parameter is selected based on the method discussed in section 3.4 with $R = 10$. Furthermore, the second tuning parameter in L_{EN} k-means method is fixed to $\alpha = 0.5$, which is a special case of the L_{EN} k-means method. Finally, each simulation is replicated 20 times.

The simulation results are shown in Tables 1-6. In particular, Table 1 shows the average clustering errors for all three methods when $p = 60$. An examination the results shows that

TABLE 1. The averaged clustering errors for various clustering methods with $p = 60$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	0.2687	0.2918	0.3258	0.3287
	L_{EN} k-means	0.2608	0.2792	0.3213	0.3253
	Standard k-means	0.2776	0.2829	0.3268	0.3300
0.6	L_1 k-means	0.2224	0.2071	0.2505	0.2942
	L_{EN} k-means	0.2021	0.2018	0.2553	0.2974
	Standard k-means	0.2158	0.2166	0.2563	0.3024
0.8	L_1 k-means	0.1087	0.1374	0.2137	0.2392
	L_{EN} k-means	0.0989	0.1342	0.1982	0.2355
	Standard k-means	0.1203	0.1474	0.2171	0.2526

TABLE 2. The averaged clustering errors for various clustering methods with $p = 80$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	0.2974	0.3145	0.3353	0.3353
	L_{EN} k-means	0.2845	0.3195	0.3342	0.3237
	Standard k-means	0.2966	0.3216	0.3429	0.3355
0.6	L_1 k-means	0.2079	0.2453	0.2661	0.3126
	L_{EN} k-means	0.2189	0.2374	0.2745	0.3066
	Standard k-means	0.2200	0.2500	0.2818	0.3139
0.8	L_1 k-means	0.1332	0.1547	0.1939	0.2616
	L_{EN} k-means	0.1011	0.1421	0.1882	0.2650
	Standard k-means	0.1174	0.1463	0.1926	0.2658

the clustering errors increase with increasing variable correlation factor ρ. This means that all of the clustering methods tend to be less effective for high-correlation conditions. It is not surprising to find that the clustering errors for all of the methods decrease with increasing vector means. We note that a larger mean provides bigger differences between the four overlapping clusters. Finally, it is important to note that the L_{EN} k-means method shows the best performance in most cases, the L_1 k-means method outperforms in other cases and the standard k-means method never outperforms the other two methods simultaneously. It is important to point out that the L_1 k-means method is a special case of L_{EN} k-means when $\alpha = 1$. Thus, the L_{EN} k-means can outperform the other methods with an appropriate value of α. Tables 2 and 3 present the results for the cases with $p = 80$ and $p = 100$, respectively. Similar results are obtained and the effect of p is not obvious.

Furthermore, Table 4 shows the average numbers of selected variables for various clustering methods with $p = 60$. It can be easily observed that the number of variables selected in the L_1 k-means method is the closest to the true number of informative variables in all of the cases. The L_{EN} k-means method has the ability to select the variables while the standard k-means method cannot perform variable selection. Similar results can be found in Tables 5 and 6.

B. CLUSTERING RESULTS WITH UNKNOWN K

In this simulation, the true number of clusters K is unknown, so we set a candidate set $K = 2, 3, 4, 5, 6$ for all clustering methods for the purpose of parameter selection. Again,

TABLE 3. The averaged clustering errors for various clustering methods with $p = 100$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	0.2924	0.2974	0.3132	0.3337
	L_{EN} k-means	0.2750	0.2816	0.3142	0.3437
	Standard k-means	0.3050	0.3087	0.3197	0.3458
0.6	L_1 k-means	0.1616	0.2332	0.2692	0.2961
	L_{EN} k-means	0.1839	0.2279	0.2618	0.3029
	Standard k-means	0.1874	0.2461	0.2645	0.3053
0.8	L_1 k-means	0.1037	0.1766	0.1924	0.2497
	L_{EN} k-means	0.1187	0.1555	0.2084	0.2579
	Standard k-means	0.1179	0.1671	0.2129	0.2595

TABLE 4. The averaged numbers of selected variables for various clustering methods with $p = 60$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	50.10	45.20	49.45	54.30
	L_{EN} k-means	59.00	58.85	58.30	59.05
	Standard k-means	60.00	60.00	60.00	60.00
0.6	L_1 k-means	53.60	54.75	51.70	50.95
	L_{EN} k-means	59.35	59.45	59.35	59.35
	Standard k-means	60.00	60.00	60.00	60.00
0.8	L_1 k-means	57.00	57.00	55.80	54.85
	L_{EN} k-means	59.65	59.55	59.20	59.65
	Standard k-means	60.00	60.00	60.00	60.00

TABLE 5. The averaged numbers of selected variables for various clustering methods with $p = 80$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	55.70	64.15	61.00	63.65
	L_{EN} k-means	76.25	77.45	76.70	77.45
	Standard k-means	80.00	80.00	80.00	80.00
0.6	L_1 k-means	66.10	66.50	68.50	63.60
	L_{EN} k-means	77.35	78.55	79.00	78.00
	Standard k-means	80.00	80.00	80.00	80.00
0.8	L_1 k-means	73.45	70.90	68.25	61.70
	L_{EN} k-means	78.65	77.40	78.95	78.55
	Standard k-means	80.00	80.00	80.00	80.00

TABLE 6. The averaged numbers of selected variables for various clustering methods with $p = 100$ (K is known).

μ	Methods	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
0.4	L_1 k-means	61.40	68.05	75.80	71.75
	L_{EN} k-means	96.20	96.85	94.55	97.65
	Standard k-means	100.00	100.00	100.00	100.00
0.6	L_1 k-means	74.15	74.20	65.80	77.20
	L_{EN} k-means	94.90	96.20	95.75	96.95
	Standard k-means	100.00	100.00	100.00	100.00
0.8	L_1 k-means	81.90	78.45	78.80	80.15
	L_{EN} k-means	98.05	98.20	95.90	96.50
	Standard k-means	100.00	100.00	100.00	100.00

for the regularized k-means methods, the candidate set for tuning parameters λ_1 and λ_2 is $\{0.05\Delta; \Delta = 1, 2, \ldots, 10\}$. The optimal parameter is selected based on the method discussed in section 3.4 with $R = 10$. Furthermore, the second tuning parameter in L_{EN} k-means method is fixed to $\alpha = 0.5$. Finally, each simulation is conducted with a replication of 20 times.

TABLE 7. The averaged clustering errors for various clustering methods with $\rho = 0.8$ (K is unknown).

μ	Methods	$p = 60$	$p = 80$	$p = 100$
0.4	L_1 k-means	0.2892	0.3029	0.2939
	L_{EN} k-means	0.2868	0.2971	0.3058
	Standard k-means	0.3016	0.3113	0.3082
0.6	L_1 k-means	0.2603	0.2621	0.2658
	L_{EN} k-means	0.2566	0.2589	0.2516
	Standard k-means	0.2595	0.2687	0.2600
0.8	L_1 k-means	0.2137	0.2103	0.1937
	L_{EN} k-means	0.2076	0.2095	0.1766
	Standard k-means	0.2147	0.2163	0.1800

TABLE 8. The averaged numbers of selected variables for various clustering methods with $\rho = 0.8$ (K is unknown).

μ	Methods	$p = 60$	$p = 80$	$p = 100$
0.4	L_1 k-means	50.05	59.95	68.45
	L_{EN} k-means	59.7	79.25	97.2
	Standard k-means	60.00	80.00	100.00
0.6	L_1 k-means	55.9	65.85	71.3
	L_{EN} k-means	59.5	79.15	97.7
	Standard k-means	60.00	80.00	100.00
0.8	L_1 k-means	56.5	68.15	81.9
	L_{EN} k-means	59.7	78.5	97.75
	Standard k-means	60.00	80.00	100.00

TABLE 9. The modified data structure in the example.

$cumstomer$	Week 1		Week 2		...	Week 54	
1	F_1^1	M_1^1	F_2^1	M_2^1	...	F_{54}^1	M_{54}^1
2	F_1^2	M_1^2	F_2^2	M_2^2	...	F_{54}^2	M_{54}^2
...	...	...	...	...	...	...	...
n	F_1^n	M_1^n	F_2^n	M_2^n	...	F_{54}^n	M_{54}^n

Without loss of generality, we only provide the results for $\rho = 0.8$. In particular, Tables 7 and 8 present the average clustering errors and the average number of variable selected for different clustering methods. Basically, the results are consistent with what found in the previous case.

C. BRIEF SUMMARY

Based on the previous simulation study, we can observe the advantages of the proposed regularized k-means methods. Although in a few cases, the L_1 k-means method outperforms the proposed L_{EN} k-mean method, we already have shown that the L_1 k-means method is merely a special case of the L_{EN} k-means method that has a much higher flexibility by introducing the second tuning parameter λ.

In addition to the clustering error, the variable selection property of the proposed method is even more important. It should be noted that if the correct informative variables are extracted, business organizations can provide more powerful promotion strategy by focusing on the important variables.

IV. CASE STUDY

In this section, we apply our proposed method in a real example. The example is regarding the online customer segmentation from a company in UK. This example was studied

TABLE 10. Clustering results for the example.

Methods	Class	Total Frequency			Total Money (pound)		
		mean	max	min	mean	max	min
L_1 k-means (98)	Class 1	60	74	96	244804.7	280206	194550.8
	Class 2	40.07317	210	21	21625.8	143825.1	1296.44
L_{EN} k-means (106)	Class 1	60	74	96	244804.7	280206	194550.8
	Class 2	40.07317	210	21	21625.8	143825.1	1296.44
Standard k-means (108)	Class 1	60	74	96	244804.7	280206	194550.8
	Class 2	40.07317	210	21	21625.8	143825.1	1296.44

by Chen *et al.* [36]. The customer transaction dataset contains 11 variables and 22190 valid transaction records in total for approximately one year. The 11 variables contain information such as invoice number, quantity, price, address, and postcode.

Instead of extracting information based on the traditional RFM model, we formulate a dataset for each customer with their weekly purchase frequency and weekly money spent and an abstract example is shown in Table 9 where F_i^j and M_i^j are the shopping frequency and money spent for customer j during week i, respectively. In the example, we focus on a small group of important customers. Thus, we keep the customer records with the purchase times larger than 20 and number of purchase weeks larger than 15. Finally, we obtain a dataset of 85 observations with 108 dimensions.

The clustering results for the three methods are shown in Table 10. We provide the summary statistics for the total frequency and money spent of each group of customers. We note that all of the methods cluster the customers into two groups, and present exactly the same clustering results. Furthermore, the numbers of selected variables by each method are shown in the brackets under their names. It is easy to discover that L_1 k-means method and L_{EN} k-mean method select 98 and 106 variables from the 108 variables, respectively. However, the standard k-means method is not able to select any variable from the whole set. Noted that the 108 variables we obtained from the dataset are the potential features which may influence the clustering results. The more features remain, the more efforts and resources are needed due to management and promotion purposes when characterizing the customers' behavior. In summary, although the proposed method and standard k-means method show no difference in terms of the clustering results, the proposed method is more efficient and effective by reducing the number of potential useful variables. Consequently, the proposed method reduces the burden from data collection and let the managers to focus on less features when making customer promotion schemes.

V. CONCLUDING REMARKS AND RESEARCH LIMITATIONS

The omnichannel business has become a hot topic due to the fast development of e-commerce and the customers' acquaintance with multichannel shopping mode. Various business organizations have started to work on the omnichannel issue in order to satisfy the new trend of customer demand. The RFM model and the k-means clustering method are typical approaches to segmentation of customers. To extend the RFM model, we divide the total frequency and monetary information into weekly level data, leading to the problem of high-dimensional analysis. To address this issue, in this paper we extend the regularized k-means clustering method with L_1-norm for independent case to the clustering method with elastic net penalty with a focus on correlated variables. Our simulation results show that the proposed method generates smaller clustering error compared with the standard k-means method, which indicates a better performance. Furthermore, our method is able to select variables during the clustering process. In particular, the number of variables is reduced from 108 to 98 without any loss on clustering accuracy in the application on real example.

This research does not discuss the tuning parameter selection for α in the L_{EN} k-means method. However, the flexibility of the L_{EN} k-means method is obtained using such parameters. This issue should be investigated in further research. We also limit our research on the accuracy of the proposed method, so the computational complexity is beyond the scope of this study. However, such issue is important for engineering practicability, which is worthy of further research. Furthermore, we limit our research to the k-means based clustering method. Other clustering methods such as EM can also be modified by adding a regularization term. In this research, we apply the proposed method to a real example, but little interpretation of the clustering results is discussed which is also important for the sense of empirical study. This can be definitely one of our future research directions.

REFERENCES

[1] C.-H. Cheng and Y.-S. Chen, "Classifying the segmentation of customer value via RFM model and RS theory," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4176–4184, Apr. 2009.

[2] C.-C.-H. Chan, C.-B. Cheng, and W.-C. Hsien, "Pricing and promotion strategies of an online shop based on customer segmentation and multiple objective decision making," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14585–14591, Nov. 2011.

[3] R.-S. Wu and P.-H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," *Electron. Commerce Res. Appl.*, vol. 10, no. 3, pp. 331–341, May 2011.

[4] M. O'Brien, Y. Liu, H. Y. Chen, and R. Lusch, "Gaining insight to B2B relationships through new segmentation approaches: Not all relationships are equal," *Expert Syst. Appl.*, vol. 161, Dec. 2020, Art. no. 113767.
[5] H. Brotspies and A. Weinstein, "Rethinking business segmentation: A conceptual model and strategic insights," *J. Strategic Marketing*, vol. 27, no. 2, pp. 164–176, Feb. 2019.
[6] F. Liu, "3D block matching algorithm in concealed image recognition and E-commerce customer segmentation," *IEEE Sensors J.*, vol. 20, no. 20, pp. 11761–11769, Aug. 2020.
[7] J. Griffin, *Customer Segmentation: Divide and Prosper. IQ Magazine*. San Jose, CA, USA: Cisco, Mar./Apr. 2003.
[8] Z. You, Y.-W. Si, D. Zhang, X. Zeng, S. C. H. Leung, and T. Li, "A decision-making framework for precision marketing," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3357–3367, May 2015.
[9] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6491–6507, Nov. 2013.
[10] A. M. Hughes, *Strategic Database Marketing*. Chicago, IL, USA: Probus Publishing Company, 1994.
[11] A. M. Hughes, "Boosting response with RFM," *Marketing Tools*, vol. 5, pp. 4–7, Jan. 1996.
[12] J. R. Bult and T. Wansbeek, "Optimal selection for direct mail," *Marketing Sci.*, vol. 14, no. 4, pp. 378–394, Nov. 1995.
[13] F. Newell, *The New Rules of Marketing: How to use one-to-one Relationship Marketing to be the Leader in Your Industry*. New York, NY, USA: McGraw-Hill, 1997.
[14] J. R. Miglautsch, "Thoughts on RFM scoring," *J. Database Marketing Customer Strategy Manage.*, vol. 8, no. 1, pp. 67–72, Aug. 2000.
[15] J. Wu and Z. Lin, "Research on customer segmentation model by clustering," in *Proc. 7th Int. Conf. Electron. Commerce (ICEC)*, Aug. 2005, pp. 316–318.
[16] J. Lee and S. Park, "Intelligent profitable customers segmentation system based on business intelligence tools," *Expert Syst. Appl.*, vol. 29, no. 1, pp. 145–152, Jul. 2005.
[17] J. Cao, Z. Bu, Y. Wang, H. Yang, J. Jiang, and H.-J. Li, "Detecting prosumer-community groups in smart grids from the multiagent perspective," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 49, no. 8, pp. 1652–1664, Aug. 2019.
[18] Z. Bu, H.-J. Li, C. Zhang, J. Cao, A. Li, and Y. Shi, "Graph K-means based on leader identification, dynamic game, and opinion dynamics," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 7, pp. 1348–1361, Jul. 2020.
[19] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, vol. 1, 1967, pp. 281–297.
[20] A. Mesforoush and M. J. Tarokh, "Customer profitability segmentation for SMEs case study: Network equipment company," *Int. J. Res. Ind. Eng.*, vol. 2, pp. 30–44, Mar. 2013.
[21] A. Amiri, "Customer-oriented catalog segmentation: Effective solution approaches," *Decis. Support Syst.*, vol. 42, no. 3, pp. 1860–1871, Dec. 2006.
[22] B. Fan and P. Z. Zhang, "Spatially enabled customer segmentation using a data classification method with uncertain predicates," *Decis. Support Syst.*, vol. 47, no. 4, pp. 343–353, 2009.
[23] A. Onan, "Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering," *IEEE Access*, vol. 7, pp. 145614–145633, 2019.
[24] A. Onan and S. Korukoğlu, "A feature selection model based on genetic rank aggregation for text sentiment classification," *J. Inf. Sci.*, vol. 43, no. 1, pp. 25–38, Feb. 2017.
[25] Y. Kim and W. N. Street, "An intelligent system for customer targeting: A data mining approach," *Decis. Support Syst.*, vol. 37, no. 2, pp. 215–228, May 2004.
[26] J.-T. Wei, M.-C. Lee, H.-K. Chen, and H.-H. Wu, "Customer relationship management in the hairdressing industry: An application of data mining techniques," *Expert Syst. Appl.*, vol. 40, no. 18, pp. 7513–7518, Dec. 2013.
[27] R. J. Kuo, J. L. Liao, and C. Tu, "Integration of ART2 neural network and genetic K-means algorithm for analyzing Web browsing paths in electronic commerce," *Decis. Support Syst.*, vol. 40, no. 2, pp. 355–374, Aug. 2005.
[28] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimension, low sample size data," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 67, no. 3, pp. 427–444, Jun. 2005.
[29] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering," *J. Amer. Stat. Assoc.*, vol. 105, no. 490, pp. 713–726, 2010.
[30] W. Sun, J. Wang, and Y. Fang, "Regularized k-means clustering of high-dimensional data and its asymptotic consistency," *Electron. J. Statist.*, vol. 6, pp. 148–167, 2012.
[31] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.
[32] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[33] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.
[34] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc., B, Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, Feb. 2006.
[35] S. Wang and J. Zhu, "Variable selection for model-based high-dimensional clustering and its application to microarray data," *Biometrics*, vol. 64, no. 2, pp. 440–448, Jun. 2008.
[36] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," *J. Database Marketing Customer Strategy Manage.*, vol. 19, no. 3, pp. 197–208, Sep. 2012.

HONG-HAO ZHAO received the bachelor's degree in industrial engineering from Tsinghua Univeristy, Beijing, China, in 2009, and the Ph.D. degree in systems engineering and engineering management from the City University of Hong Kong, Hong Kong, China, in 2014.

He is currently an Associate Professor with the Department of Decision Sciences, School of Business, Macau University of Science and Technology, Taipa, Macau. His research interests include data mining, machine learning, business data analysis, and statistical process control.

XI-CHUN LUO received the Master of Charity and Philanthropy Management degree from the Macau University of Science and Technology, Taipa, China, in 2017. His research interests include data mining and game theory.

RUI MA received the bachelor's degree in software engineering from Jilin University, Jilin, China, in 2009, and the master's degree in business information systems from the City University of Hong Kong, Hong Kong, China, in 2013. She is currently pursuing the Ph.D. degree in management with the Macau University of Science and Technology, Taipa, China.

From 2013 to 2015, she was a Research Assistant with the City University of Hong Kong. She is currently a Lecture with the Department of Computer Science, Zhuhai College of Jilin University. Her research interests include data mining and machine learning.

XI LU received the bachelor's degree in electronic information engineering from the University of Electronic Science and Technology of China, Sichuan, China, in 2018, and the master's degree in business analytics from the Macau University of Science and Technology, Taipa, China, 2020.