Sparse Regularization in Fuzzy c-Means for High-Dimensional Data Clustering

Xiangyu Chang, Qingnan Wang, Yuewen Liu, and Yu Wang

***Abstract*—In high-dimensional data clustering practices, the cluster structure is commonly assumed to be confined to a limited number of relevant features, rather than the entire feature set. However, for high-dimensional data, identifying the relevant features and discovering the cluster structure are still challenging problems. To solve these problems, this paper proposes a novel fuzzy c-means (FCM) model with sparse regularization ($\ell_q(0 < q \leq 1)$-norm regularization), by reformulating the FCM objective function into the weighted between-cluster sum of square form and imposing the sparse regularization on the weights. An algorithm is also developed to explicitly solve the proposed model. Compared with the existing clustering models, the proposed model can shrink the weights of irrelevant features (noisy features) to exact zero, and also can be efficiently solved in analytic forms when $q = 1, 1/2$. Experiments on both synthetic and real-world data sets show that the proposed approach outperforms the existing clustering approaches.**

***Index Terms*—$\ell_q(0 < q \leq 1)$-norm regularization, fuzzy c-means (FCM), high-dimensional data clustering.**

I. Introduction

HIGH-DIMENSIONAL data clustering problems, i.e., the objects to be clustered have a large number of features, are still challenging problems in recent years [1]–[4]. In most of the real-world cases, only a small portion of the features is assumed to be relevant to the cluster structure [1], [2], [4]. For example, only a tiny portion of genes (*relevant features*) are responsible for a certain biological activity, while the others are irrelevant (*noisy features*). A good clustering approach should be able to identify the relevant features and avoid the negative influences of the noisy features [1], [2]. Intuitively, if we can assign positive weights to the relevant features and assign exact zero weights to the noisy features, the negative influences of the noisy features could be avoided.

Manuscript received June 1, 2016; revised September 10, 2016; accepted October 28, 2016. This work was supported in part by the National Science Foundation of China under Grant 11401462, Grant 61502342, Grant 71301128, and Grant 71331005, and in part by the China Post-Doctoral Science Foundation under Grant 2015M582630, Grant 2014M560795, and Grant 2015T81039. This paper was recommended by Associate Editor A. F. Skarmeta Gomez. *(Corresponding author: Yuewen Liu.)*

X. Chang, Q. Wang, and Y. Liu are with the Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: liuyuewen@mail.xjtu.edu.cn).

Y. Wang is with the Department of Statistics, University of California at Berkeley, Berkeley, CA 94720 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2016.2627686

Fuzzy c-means (FCM), as one of the classical fuzzy type clustering approaches, has been well studied and extended in [5]–[14]. Based on FCM, several approaches are developed to address the high-dimensional clustering problems [2], [8]–[10], [13], [15]. One stream of the approaches is to reduce the data dimensions before clustering the objects. The dimension reduction methods, include *principle components analysis* [16] and *non-negative matrix factorization* [17]. However, evidences show that the principal components do not provide reasonable representatives of the original dimensions [16]. Another stream of approaches, named as *attribute-weighting approaches*, is to weight the attributes differently. These approaches equip a weight vector into the FCM objective function to indicate the relevance of features [8]–[10], [15]. However, none of these attribute-weighting approaches can shrink the weights of noisy features to exact zero, thus a considerable proportion of noisy features may be remained and negatively affect the clustering results. These drawbacks significantly obstruct the application of the FCMs on high-dimensional data clustering problems.

Therefore, a natural research question arises: can the FCM be extended to shrink the weights of noisy features to exact zero? Such an extension may dramatically reduce the negative influences of noisy features and improve the clustering performance in solving the high-dimensional data clustering problems. Unfortunately, this research question has not been answered yet.

To address this research question, we try to impose sparse regularizations on FCM. We first justify that FCM can be reformulated into Witten and Tibshirani's [1] *sparse clustering framework*. Witten and Tibshirani's [1] sparse clustering framework offered a specific attribute-weighting method, which optimizes a weighted cost objective function using the ℓ_1-norm regularization technique, thus is able to assign exact zero weights to noisy features [1]. Based on the justification, we propose a sparse FCM which maximizes the *weighted between-cluster sum of squares* (BCSS) with ℓ_1-norm regularization. Furthermore, inspired by the broad literature which shows the outperformance of the nonconvex $\ell_q(0 < q < 1)$-norm in data mining problems [18]–[23], we extend the proposed model to a sparse FCM with $\ell_q(0 < q \leq 1)$-norm regularization, in the purpose of utilizing the outperformance of $\ell_q(0 < q < 1)$-norm.

The major contributions of this paper are summarized as follows.

1) The proposed novel sparse FCM model imposes the sparse regularization ($\ell_{0<q\leq 1}$-norm regularization) to

TABLE I
NOTATIONS USED IN THIS PAPER

Notations	Description
$\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$	Data set in the matrix form
$\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{X}_j \in \mathbb{R}^n$	The ith row and jth column of $\mathbf{X}$
$\mathbf{C} = (c_{kj}) \in \mathbb{R}^{K\times p}$	Cluster centers
$C_k, k = 1, \ldots, K$	The kth cluster center
$U = (u_{ik})^\top \in \mathbb{R}^{n\times K}$	Membership Degrees
$\mathcal{G} = \{G_1, G_2, \ldots, G_K\}$	K groups
$\mathbf{w} = (w_1, \ldots, w_p)^\top$	Feature weights
$a_j = BCSS(\mathcal{G})_j$	BCSS of the jth feature
n	The number of observations
α	Fuzzy index of fuzzy memberships
β	Fuzzy index of fuzzy weighting
p	The number of features

assign exact zero weights to the noisy features for clustering high-dimensional data. The proposed model is motivated by the sparse clustering framework [1], but extends the framework.

2) An algorithm is subtly designed to solve the nonsmooth and nonconvex optimization problem of the proposed model. When $q = 1, 1/2$, the proposed algorithm has closed form solutions in all the steps. These closed form solutions make the proposed approach more tractable and efficient in dealing with high-dimensional data sets.
3) The outperformance of the proposed approach comparing to some related clustering approaches is demonstrated by expensive experiments on both synthetic and real-world data sets.

The remainder of this paper is organized as follows. Section II introduces the proposed FCM with sparse regularization, and investigates an efficient algorithm for pursuing the solution. Section III discusses the related work in the literature. Section IV illustrates and compares the finite sample performance of the proposed FCM with the related clustering approaches using both synthetic and real-world data sets. Section V concludes this paper with discussions. All the theoretical proofs are relegated to Appendix. For clarity, the notations used in this paper are defined in Table I.

II. SPARSE REGULARIZATION IN FUZZY C-MEANS

A. Notations

Let $\mathbf{X} = (x_{ij}) \in \mathbb{R}^{n\times p}$ denote a data set in the matrix form with n objects and p features. We assume that $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{X}_j \in \mathbb{R}^n$ are the ith row and jth column of $\mathbf{X}$, respectively, then $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p] = [\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top]^\top$. It is well known that the standard FCM clusters the data into K groups $\mathcal{G} = \{G_1, G_2, \ldots, G_K\}$ by minimizing the sum of distances between the objects and the corresponding cluster centers $\mathbf{C} = (C_1^\top, C_2^\top, \ldots, C_K^\top)^\top = (c_{kj}) \in \mathbb{R}^{K\times p}$, where $C_k, k = 1, \ldots, K$ is the kth cluster center. Then FCM can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\mathcal{G},\mathbf{C}} \ & \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}^{\alpha} d(\mathbf{x}_i, C_k) \\ \text{s.t.} \ & \sum_{k=1}^{K} u_{ik} = 1, 0 \le u_{ik} \le 1 \\ & i = 1, \ldots, n, k = 1, \ldots, K \end{aligned} \tag{1}$$

where $U = (u_{ik})^\top \in \mathbb{R}^{n\times K}$, u_{ik} denotes the degree of membership of the ith object belonging to the kth fuzzy cluster, $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a dissimilarity measure satisfying $d(a, a) = 0, d(a, b) \ge 0$ and $d(a, b) = d(b, a)$ and $\alpha \ge 1$ denotes the weighting exponent that controls the extent of membership sharing between fuzzy clusters. As for the dissimilarity between vector $\mathbf{x}_i$ and $\mathbf{x}_j$, it is a common practice to use the square of Euclidean distance (also called ℓ_2-norm), namely, $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \sum_{l=1}^{p}(x_{il} - x_{jl})^2$.

A high-dimensional data clustering problem commonly assumes that the data to be clustered have a large number of noisy features, and the cluster structure is confined to a limited number of relevant features rather than the entire feature set [1], [3], [15]. The standard FCM cannot deal with such high-dimensional data, because it cannot select relevant features and identify the cluster structure simultaneously. To overcome this hurdle, we propose a novel sparse FCM approach in the next section.

B. Sparse Fuzzy c-Means

Witten and Tibshirani [1] verified that the classical k-means and hierarchical clustering models can be reformulated using the following framework:

$$\max_{\Theta(\mathcal{G})} \sum_{j=1}^{p} f_j\big(\mathbf{X}_j, \Theta(\mathcal{G})\big) \tag{2}$$

where $f_j(\mathbf{X}_j, \Theta(\mathcal{G}))$ is a function related only to the jth feature of the data, and $\Theta(\mathcal{G})$ is the model parameter. They further defined a *sparse clustering framework*

$$\begin{aligned} \max_{\mathbf{w},\Theta(\mathcal{G})} \ & \sum_{j=1}^{p} w_j f_j\big(\mathbf{X}_j, \Theta(\mathcal{G})\big) \\ \text{s.t.} \ & \|\mathbf{w}\|_2 \le 1, \|\mathbf{w}\|_1 \le s, w_j \ge 0 \end{aligned} \tag{3}$$

where s is a tunning parameter and $\|\mathbf{w}\|_1 = \sum_{j=1}^{p}|w_j|$ is the ℓ_1-norm of $\mathbf{w}$. Here, w_j can be interpreted as the contribution of the jth feature to the objective function (3). The ℓ_1-norm has been proved to be able to generate a sparse solution in many applications, where the tuning parameter s controls the number of relevant features in the clustering results [1], [24]. Next, we provide Lemma 1 which justifies that the FCM is also a special case of the framework (2).

Lemma 1: Suppose $\mathbf{X}$ is the data matrix, then

$$\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}^{\alpha}\|\mathbf{x}_i - C_k\|^2 = \sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n} u_{ik}^{\alpha}u_{i'k}^{\alpha}\|\mathbf{x}_i - \mathbf{x}_{i'}\|^2$$

where $n_k = \sum_{i=1}^{n} u_{ik}^{\alpha}$, and C_k is selected as a *weighted empirical mean*, which is

$$C_k = \frac{\sum_{i=1}^{n} u_{ik}^{\alpha}\mathbf{x}_i}{n_k}. \tag{4}$$

Proof: We know that

$$\begin{aligned}\sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2&=\sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\|\mathbf{x}_i-C_k+C_k-\mathbf{x}_{i'}\|^2\\&=\sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\Big\{\|\mathbf{x}_i-C_k\|^2\\&\qquad+\|\mathbf{x}_{i'}-C_k\|^2+2(\mathbf{x}_i-C_k)^{\top}(\mathbf{x}_i-C_k)\Big\}.\end{aligned}$$

Since $n_k=\sum_{i'=1}^{n}u_{i'k}^{\alpha}$, we have

$$\begin{aligned}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\|\mathbf{x}_i-C_k\|^2&=\sum_{i'=1}^{n}u_{i'k}^{\alpha}\sum_{i=1}^{n}u_{ik}^{\alpha}\|\mathbf{x}_i-C_k\|^2\\&=n_k\sum_{i=1}^{n}u_{ik}^{\alpha}\|\mathbf{x}_i-C_k\|^2\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}(\mathbf{x}_i-C_k)^{\top}(\mathbf{x}_i-C_k)\\&\quad=\left[\sum_{i=1}^{n}u_{ik}^{\alpha}(\mathbf{x}_i-C_k)\right]\left[\sum_{i'=1}^{n}u_{i'k}^{\alpha}(\mathbf{x}_{i'}-C_k)\right]=0.\end{aligned}$$

Then

$$\sum_{k=1}^{K}\frac{1}{2n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\|\mathbf{x}_i-\mathbf{x}_{i'}\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}u_{ik}^{\alpha}\|\mathbf{x}_i-C_k\|^2.$$

■

Intuitively, the left hand side of Lemma 1 is the objective function (1) of FCM, while the right hand side of Lemma 1 evaluates the dissimilarity within cluster, which is referred as *within-cluster sum of square* of FCM. In fact, an operational definition of clustering can be stated as follows: given a representation of n objects, finding K groups based on a measure of dissimilarity such that objects within the same group are alike but objects in different groups are disparate [25]. Therefore, we can also model the BCSS of FCM (1) as

$$\begin{aligned}\mathrm{BCSS}(\mathcal{G})_j=\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n}\big(x_{ij}-x_{i'j}\big)^2\\-\sum_{k=1}^{K}\frac{1}{n_k}\sum_{i=1}^{n}\sum_{i'=1}^{n}u_{ik}^{\alpha}u_{i'k}^{\alpha}\big(x_{ij}-x_{i'j}\big)^2\end{aligned}\tag{5}$$

where $\mathrm{BCSS}(\mathcal{G})=(\mathrm{BCSS}(\mathcal{G})_1,\ldots,\mathrm{BCSS}(\mathcal{G})_p)^{\top}$. Then, we are able to rewrite FCM as

$$\begin{aligned}\max_{\mathcal{G},\mathbf{C}}\ &\sum_{j=1}^{p}\mathrm{BCSS}(\mathcal{G})_j\\\text{s.t.}\ &\sum_{k=1}^{K}u_{ik}=1,0\le u_{ik}\le 1\\&i=1,\ldots,n,k=1,\ldots,K.\end{aligned}\tag{6}$$

Note that $\mathrm{BCSS}(\mathcal{G})_j, j=1,\ldots,p$ is a function which is only related to the jth feature. In other words, FCM satisfies the framework (2). According to Witten and Tibshirani's [1] sparse clustering framework, the FCM can be generalized to the following model:

$$\begin{aligned}\max_{\mathcal{G},\boldsymbol{C},\mathbf{w}}\ &F(U,\mathbf{C},\mathbf{w})=\mathbf{w}^{\top}\mathrm{BCSS}(\mathcal{G})\\\text{s.t.}\ &\sum_{k=1}^{K}u_{ik}=1,0\le u_{ik}\le 1\\&\|\mathbf{w}\|_2\le 1,\|\mathbf{w}\|_q^q\le s\\&w_j\ge 0,j=1,\ldots,p\end{aligned}\tag{7}$$

where $0<q\le 1$ and $\|\mathbf{w}\|_q^q=\sum_{j=1}^{p}|w_j|^q$. We call (7) as the sparse FCM model. Note that Witten and Tibshirani's [1] sparse clustering framework only considers the case $q=1$. However, there are also some evidences showing the outperformance of the nonconvex $\ell_q(0<q<1)$-norm regularization in SVM [18], compressive sensing [19], SAR imaging recovery [20], robust regression [21], matrix completion [22], and penalized clustering [23]. In order to utilize the outperformance of the $\ell_q(0<q<1)$-norm regularization, we generalize the standard ℓ_1-norm regularization to the $\ell_q(0<q\le 1)$-norm regularization. Nevertheless, this generalization leads to a nonsmooth and nonconvex optimization problem (e.g., $0<q<1$). The following section focuses on solving the optimization problem.

C. Algorithm

For the simplicity, denote $a_j=\mathrm{BCSS}(\mathcal{G})_j$, $\mathbf{a}=(a_1,\ldots,a_p)^{\top}$ and the objective function of (7) as $F(U,\mathbf{C},\mathbf{w})=\sum_{j=1}^{p}w_ja_j$. We apply the alternative iteration technique to construct an algorithm to solve the model (7). We first fix $\mathbf{C}$ and $\mathbf{w}$ and maximize $F(U)$ with respect to U, and then we fix $\mathbf{w}$ and U and maximize $F(\mathbf{C})$ with respect to $\mathbf{C}$. Finally, we fix U and $\mathbf{C}$ and maximize $F(\mathbf{w})$ with respect to $\mathbf{w}$. To this end, the following Theorems 1 and 2 show the detailed calculations.

Theorem 1: Let the cluster centers $\mathbf{C}$ and the attribute weights $\mathbf{w}$ be fixed, $F(U)$ is minimized if

$$u_{ik}=\begin{cases}\dfrac{1}{P_k}\\\text{if } D_{ik}=0 \text{ and } P_k=\mathrm{Card}\{j:D_{ik}=0\}\\0\\\text{if } D_{ik}\ne 0 \text{ but } D_{\mathrm{it}}=0 \text{ for some } t,\ t\ne k\\\dfrac{1}{\sum_{t=1}^{K}\left(\frac{D_{ik}}{D_{tk}}\right)^{\left(\frac{1}{\alpha-1}\right)}}\ \text{otherwise}\end{cases}\tag{8}$$

where $D_{ik}=\sum_{k=1}^{K}w_j(x_{ij}-c_{kj})^2$ and $\mathrm{Card}(A)$ is the cardinality of set A.

Theorem 2: Let $\mathbf{w}$ and U be fixed, and $F(\mathbf{C})$ is minimized if

$$c_{kj}=\begin{cases}0 & \text{if } w_j=0\\\dfrac{\sum_{i=1}^{n}u_{ik}^{\alpha}x_{ij}}{\sum_{i=1}^{n}u_{ik}^{\alpha}} & \text{if } w_j\ne 0\end{cases}\tag{9}$$

where $k=1,\ldots,K,j=1,\ldots,p$.

Algorithm 1 FCM With $\ell_q(0 < q \leq 1)$-Norm Regularization

Input:

The number of clusters K and data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$.

Output:

Clusters $C_1, C_2, \ldots, C_K$ and $\mathbf{w}^{old}$.

1: Initialize $\mathbf{w}$ as $w_1^{old} = w_2^{old} = \cdots = w_p^{old} = \frac{1}{\sqrt{p}}$;

2: Update the partition matrix U by (8);

3: Update the cluster centers $\mathbf{C}$ by (9);

4: Fix $C_1, C_2, \ldots, C_K$ and U and calculate a_j. Solve the optimization problem

$$\max_{w} \sum_{j=1}^{p} w_j a_j. \quad \text{s.t.} \quad \|\mathbf{w}\|_2^2 \leq 1, \|\mathbf{w}\|_q^q \leq s \tag{10}$$

to get $\mathbf{w}^{new}$;

5: Repeat the steps 2, 3 and 4 until the stopping criterion is satisfied

$$\frac{\sum_{j=1}^{p} |w_j^{new} - w_j^{old}|}{\sum_{j=1}^{p} |w_j^{old}|} < 10^{-4}.$$

Theorems 1 and 2 indicate the optimal solutions of degree memberships U and cluster centers $\mathbf{C}$, respectively. According to Theorems 1 and 2, a detailed algorithm for solving (7) is illustrated below.

A critical point of Algorithm 1 is to solve the problem (10). It is obvious that the problem (10) is a nonsmooth and nonconvex optimization problem for $0 < q < 1$, thus is difficult to be solved. To this end, we need to justify the following Theorems 3 and 4.

Without loss of generality, we assume that the sequence $\{a_j\}_{j=1}^p$ in step 4 of Algorithm 1 is ordered decreasingly, i.e., $a_i \geq a_j$ for any $i < j$. Then we have the following.

Theorem 3: If $(\|\mathbf{a}\|_q^q/\|\mathbf{a}\|_2^q) \geq s \geq 1$ and $0 < q \leq 1$, the optimal solution $\mathbf{w}^*$ of problem (10) satisfies $\|\mathbf{w}^*\|_q^q = s$ and $\|\mathbf{w}^*\|_2^2 = 1$.

Proof: Let $a_j = (1/n)\sum_{i=1}^{n}\sum_{i'=1}^{n}(x_{ij} - x_{i'j})^2 - \sum_{k=1}^{K}(1/n_k)\sum_{i=1}^{n}\sum_{i'=1}^{n} u_{ik}^{\alpha} u_{i'k}^{\alpha}(x_{ij} - x_{i'j})^2$ and $\lambda_j^* = w_j^{*q}$. Then, we need to justify $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^*)^{\intercal}$ is the optimal solution of

$$\max_{\boldsymbol{\lambda}} \sum_{j=1}^{p} \lambda_j^{1/q} a_j \quad \text{s.t.} \sum_{j=1}^{p} \lambda_j^{2/q} \leq 1, \sum_{j=1}^{p} \lambda_j \leq s$$

and satisfies $\sum_{j=1}^{p} \lambda_j^{*2/q} = 1$ and $\sum_{j=1}^{p} \lambda_j^* = s$.

If $\sum_{j=1}^{p} \lambda_j^{*2/q} < 1$, there exists a small $\delta > 0$, such that $\|\bar{\boldsymbol{\lambda}}\|_1 = \|\boldsymbol{\lambda}^*\|_1 \leq s$ where $\bar{\boldsymbol{\lambda}} = (\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_p)^{\intercal}$, $\bar{\lambda}_1 = \lambda_1^* + \delta, \bar{\lambda}_2 = \lambda_2^* - \delta, \bar{\lambda}_j = \lambda_j^*, j = 3, \ldots, p$ and $\sum_{j=1}^{p} \bar{\lambda}_j < 1$. Since $\lambda_1^* \geq \lambda_2^*$ and $f(x) = x^{1/q}$ is convex and increasing, we can get $\sum_{j=1}^{p} a_j u_{ik}^{\alpha} \bar{\lambda}_j^{1/q} > \sum_{j=1}^{p} a_j u_{ik}^{\alpha} \lambda_j^{*1/q}$, which comes to the contradiction. So the optimal condition is $\sum_{j=1}^{p} \lambda_j^{*2/q} = 1$.

If $\sum_{j=1}^{p} \lambda_j^{*2/q} = 1$ and $\sum_{j=1}^{p} \lambda_j^* < s$, $\lambda^{1/q}$ must be the optimal solution of

$$\max_{\mathbf{w}} \mathbf{w}^{\intercal}\boldsymbol{a} \quad \text{s.t.} \quad \|\mathbf{w}\|_2^2 = 1, w_j \geq 0.$$

The solution of (10) is $w_j^* = -(a_j/\|\boldsymbol{a}\|_2)$. Then we get

$$\left\|\mathbf{w}^*\right\|_q^q = \frac{\|\boldsymbol{a}\|_q^q}{\|\boldsymbol{a}\|_2^q} \geq s \tag{11}$$

which means $\|\boldsymbol{\lambda}^*\|_1 \geq s$. This also makes a contradiction. So the optimal condition satisfies $\sum_{j=1}^{p} \lambda_j^* = 1$. ■

Theorem 3 proves that the optimal solution of (10) satisfies the boundary constraints. Based on Theorem 3, we can prove that the optimal $\mathbf{w}^*$ is the real roots of an algebraic equation (12) for $0 < q < 1$, as shown in the following Theorem 4.

Theorem 4: Under the same condition of Theorem 3, there exists a constant Δ satisfies $0 \leq \Delta \leq 2\sum_{l=1}^{p} a_l^{2-q}$ such that the optimal solution of (10) with $0 < q < 1$ has the following form:

$$w_j^* = \begin{cases} F_{1j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 0 \leq \Delta < 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \\ F_{2j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \leq \Delta < 2\sum_{l=j}^{p+1} a_l^{2-q} \\ 0 \quad \text{if } \Delta \geq 2\sum_{l=j}^{p+1} a_l^{2-q} \end{cases}$$

where $j = 1, \ldots, p$, r is the scaling factor to ensure $\|\mathbf{w}^*\|_2 = 1$, and $F_{\gamma j}(\phi)$ is the γth largest real root of equation

$$a_j x_j^{1-q} - 2x_j^{2-q} - \left(\frac{1-q}{2(2-q)}\right)^{1-q} \frac{1}{2-q}\phi = 0 \tag{12}$$

where $\gamma = 1, 2$ and

$$\phi(\Delta) = \begin{cases} \Delta - T & \text{if } T \leq \Delta \leq T + a_j^{2-q} \\ T - \Delta & \text{if } T + a_j^{2-q} \leq \Delta \leq T \end{cases} \tag{13}$$

where $T = 2\sum_{h=j+1}^{p+1} a_h^{2-q}$ and $h = p, p-1, \ldots, 1$.

Intuitively, Theorem 3 shows that an appropriate Δ should satisfy $\|\mathbf{w}^*\|_q^q = s$, while Theorem 4 reveals that there will be several Δs which satisfy the constraint. In Algorithm 1, we determine to choose the Δ with the largest objective function value. The numerical methods, such as Dichotomy searching, can be employed to find the proper Δ because $\|\mathbf{w}^*\|_q^q$ is a continuous function of Δ.

Using numerical methods to solve the problem (10) is associated with a high computational complexity. To avoid the computational complexity, we further justify that the optimal solution of problem (10) has a closed form for $q = 1, 1/2$, respectively. Based on Theorems 3 and 4, we have the following.

Corollary 1: Under the same condition of Theorem 3, there exists a constant $\Delta > 0$ such that the solution of (10) for $q = 1$ is $\mathbf{w}^* = (S(\mathbf{a}, \Delta)/\|S(\mathbf{a}, \Delta)\|_2)$, where $\mathbf{a} = (a_1, \ldots, a_p)^{\top}$ and $S(\mathbf{a}, \Delta) = \max(\mathbf{a} - \Delta, 0)$.

Corollary 2: When $q = 1/2$, there exists a constant $\Delta \geq 0$ such that the optimal solution of (10) has the same form as in Theorem 4, where $F_{\gamma i}$ is defined as follows:

$$F_{\gamma i}(\phi) = \frac{2}{3} a_i \cos^2\left(\frac{\pi}{3} + (-1)^{\gamma}\frac{1}{3}\arccos\frac{\phi}{a_i^{3/2}}\right) \tag{14}$$

where $\gamma = 1, 2$.

Corollaries 1 and 2 give the closed form solutions of (10) for $q = 1, 1/2$, respectively, thus can dramatically reduce the computational complexity of Algorithm 1. In deed, Corollaries 1 and 2 also indicate that the solutions of (10) are the popular normalized soft [26] and normalized half [19] thresholding functions in sparse modeling. Here, we name the FCM with ℓ_1-norm regularization as ℓ_1-*c*-means and FCM with $\ell_{1/2}$-norm regularization as $\ell_{1/2}$-*c*-means.

Another problem in Algorithm 1 is to choose the tuning parameter s and fuzzification parameter α. For the tuning parameter s, we know that

$$p^{(1/2-1/q)}\|\mathbf{w}\|_q \leq \|\mathbf{w}\|_2 \leq \|\mathbf{w}\|_q, \quad \forall \mathbf{w} \in \mathbb{R}^p,\ 0 < q \leq 1.$$

From the inequality, we can infer that a meaningful s should be in $[1, p^{(1/q-1/2)}]$. If $s > p^{(1/q-1/2)}$, the $\ell_q(0 < q \leq 1)$-norm regularization term will be inactive, the constraints of problem (10) become $\|\mathbf{w}\|^2 \leq 1, w_j \geq 0$, and its optimal solution is $\mathbf{w}^* = (\mathbf{a}/\|\mathbf{a}\|_2)$, which is not sparse at all. Similarly, if $0 < s < 1$, the ℓ_2-norm regularization should be inactive, and constraints of problem (10) becomes $\|\mathbf{w}\|_q^q \leq s, w_j \geq 0$. The optimal solution of $\mathbf{w}$ is trivial with only one nonzero element. In conclusion, the tuning parameter s should be selected in $[1, p^{(1/q-1/2)}]$. For the fuzzification parameter α, its selection is still an open question even for FCM. According to [8], α is commonly selected from a fixed set such as $\alpha \in \{1.1, 1.2, \ldots, 3\}$. Following a procedure similar to [1], we apply a permutation technique and calculate the gap statistic [27] to select s and α simultaneously.

We should mention that although the generated iterative series in Algorithm 1 is not guaranteed to converge to the global optimum, the objective function will increase monotonically and achieve the local maximal value. Following the same assumption in [28], suppose the partition matrix U is a binary matrix, then the data matrix $\mathbf{X}$ can only have a finite number of possible partitions. Moreover, since we know the optimal weights $\mathbf{w}^*$ for each fixed partition is unique based on the subsequent analysis, it shows that the feasible set of the optimization is finite. Therefore, the algorithm will terminate after finite iterations and reach a local maximum.

Furthermore, the updating of U and $\mathbf{C}$ in Algorithm 1 can be interpreted as applying FCM on a weighted data matrix, thus the computational complexity of the steps 2 and 3 is the same as the standard FCM with $O(npK^2)$ [14]. To update $\mathbf{w}$ in step 4, we have to solve (10) by the Dichotomy searching scheme numerically. The main computational complexity of solving (10) is $O(p \log 1/\epsilon)$ where ϵ is the required error for searching. Therefore, the computational complexity of the proposed sparse FCM is $O(npK^2) + O(p \log 1/\epsilon)$.

In summary, the implementation of Algorithm 1 is as follows: update the partition matrix U by Theorem 1; update the cluster centers $\mathbf{C}$ by Theorem 2; update $\mathbf{w}$ by Corollaries 1 and 2 for $q = 1, 1/2$, respectively; and choose the tuning parameter s and fuzzification parameter α using the gap statistic.

III. Related Work

This paper is naturally related to the high-dimensional data clustering literature. High-dimensional data clustering problems have been initially studied using the strategy of reducing dimensions before clustering [16], [17]. However, some evidences show that the reduced dimensions do not provide a reasonable representative of the original dimensions [16]. To identify the most relevant dimensions, *attributed-weighting approaches* have been proposed [8]–[10], [29]–[32]. According to an up-to-date comprehensive review [2], the attributed-weighting approaches are also the major branch of the *soft subspace clustering approaches*.

Among the attributed-weighting approaches, one type of approaches is the attribute-weighting FCM (AWFCM) [8]. Assume w_j is the weight of jth feature, then AWFCM is

$$\begin{aligned}\min_{\mathcal{G},\mathbf{C},\mathbf{w}}\ & \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}^{\alpha} w_j^{\beta} \sum_{j=1}^{p}\left(x_{ij} - c_{kj}\right)^2 \\ \text{s.t.}\ & \sum_{k=1}^{K} u_{ik} = 1, 0 \leq u_{ik} \leq 1 \\ & i = 1, \ldots, n, k = 1, \ldots, K \\ & \sum_{j=1}^{p} w_j = 1, 0 \leq w_j \leq 1, j = 1, \ldots, p. \end{aligned} \tag{15}$$

In this model, the weights indicate the relevance of different features in a cluster. When $\alpha = \beta = 1$, the AWFCM degenerates to the *feature weight self-adjustment model* [32]; when $\alpha = 1$ and $\beta > 1$, AWFCM becomes the *W-k-means model* [30]. In the classification framework of soft subspace clustering approaches [2], the above AWFCM-type models are also called the *conventional soft subspace clustering* models.

A second type of approaches is the *FCM* with weight entropy regularization (WEFCM). Inspired by the clustering objects on subsets of attributes (COSA) method [29], Zhou and Chen [15] proposed the WEFCM as follows:

$$\begin{aligned}\min_{\mathcal{G},\mathbf{C},\mathbf{w}}\ & \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}^{\alpha} \sum_{j=1}^{p} w_{jk}\left(x_{ij} - c_{kj}\right)^2 \\ & + \lambda \sum_{j=1}^{p}\sum_{k=1}^{K} w_{jk} \log w_{jk} \end{aligned} \tag{16}$$

$$\begin{aligned}\text{s.t.}\ & \sum_{k=1}^{K} u_{ik} = 1, 0 \leq u_{ik} \leq 1 \\ & i = 1, \ldots, n, k = 1, \ldots, K \\ & \sum_{j=1}^{p} w_{jk} = 1, 0 < w_{jk} \leq 1 \end{aligned} \tag{17}$$

where $\lambda > 0$ is a tuning parameter. Imposing the entropy regularization can improve the interpretability of weights [29], [34]. Specially, when $\alpha = 1$, the

WEFCM reduces to the entropy weighting k-means (EWKM) model [31]. Since the weights of COSA, WEFCM, and EWKM are different for different clusters, they are classified as the *independent soft subspace clustering* in [2].

It is easy to prove that the attribute weights of the aforementioned approaches cannot be shrunk to exact zero: the weights in WEFCM (16) cannot be zero due to the logarithm function; Keller and Klawonn [8] showed that the solution of AWFCM (15) is not exact sparse, i.e., no attribute weight will be zero (also see empirical evidences in Section IV). In other words, for AWFCM and WEFCM, there are still a large portion of noisy features remained to negatively influence the clustering results. Comparing to AWFCM and WEFCM, the proposed sparse FCM (7) can shrink the weights of noisy features to exact zero (see Corollaries 1 and 2). Since AWFCM and WEFCM are representatives of related fuzzy subspace clustering models, the following experimental study will compare these approaches with our proposed approaches (i.e., the ℓ_1- and $\ell_{1/2}$-c-means).

This paper is also related to Witten and Tibshirani's [1] sparse clustering framework. Witten and Tibshirani [1] proposed a framework of sparse clustering, inspired by the resulting sparse solution of ℓ_1-norm regularization [1], [19], [24]. Their framework optimizes a weighted cost objective function using the ℓ_1-norm regularization technique, thus is able to assign exact zero weights to noisy features [1]. Witten and Tibshirani [1] justified that several traditional hard clustering models, such as k-means and hierarchical clustering, can be reformulated using their framework. Based on their justification, they developed the sparse k-means and hierarchical clustering to solve the high-dimensional data clustering problems [1]. When the sparse k-means and hierarchical clustering approaches are both hard clustering approaches, the proposed approach in this paper adapts the sparse clustering framework [1] to the FCM model. From this perspective, the proposed approach extends Witten and Tibshirani's work [1]. The following experimental study will also compare the sparse k-means (i.e., ℓ_1-k-means) with our proposed approaches.

IV. Experimental Study

In this section, we evaluate and compare the finite sample performance of ℓ_1- and $\ell_{1/2}$-c-means with FCM [34], WEFCM [15], AWFCM [8], k-means [35], and ℓ_1-k-means [1]. We consider several criteria to obtain comprehensive comparisons. The first criterion is the classification error rate (CER) [1], which is defined as $CER \triangleq \sum_{i>i'} |1_{\hat{\mathcal{G}}(i,i')} - 1_{\mathcal{G}(i,i')}|/\binom{n}{2}$, where $1_{\mathcal{G}(i,j)}$ is an indicator function to record whether the ith and jth objects are in the same group with respect to partition $\mathcal{G}$. The second criterion is the Davies Bouldin (DB) index [36], which is defined as $\mathrm{DB} = (1/K)\sum_{k=1}^{K} \max_{s,s\neq k}(S_k + S_s)/d_{ks}$, where d_{ks} is the distance between the centroid of group G_k and G_s, and $S_k = \sqrt{(1/n)\sum_{i\in G_k}\sum_{j=1}^{m}|x_{ij} - z_{kj}|^2}$ is a dispersion measure of group G_k. The third criterion is the *Dunn index* [37], which is the ratio of the smallest distance between objects who are not in the same cluster to the largest intracluster distance. The Dunn index is defined as $\mathrm{DI} = (\min_{1\leq k\leq l\leq K}\delta(G_k, G_l)/\max_{1\leq m\leq K}\Delta_m)$, where $\delta(G_k, G_l)$ is the intercluster distance between clusters G_k and G_l, and Δ_m calculates the maximum distance or the mean distance between all items within cluster G_m. The fourth criterion is the number of *nonzero weights* $\mathrm{NW} = \mathrm{Card}\{i : \hat{w}_i \neq 0\}$, where $\hat{\mathbf{w}}$ is the estimation of ground truth $\mathbf{w}$ produced by some computational algorithm. It counts the number of features selected as relevant features. The fifth criterion is the number of *proper zero weights* $\mathrm{PZW} = \mathrm{Card}\{i : w_i = 0, \hat{w}_i = 0\}$, which records the number of noisy features correctly eliminated. The sixth criterion is the number of *proper nonzero weights* $\mathrm{PNW} = \mathrm{Card}\{i : w_i \neq 0, \hat{w}_i \neq 0\}$, which measures the number of relevant features correctly selected. Note that PZW and PNW are only available when the true relevant features are known, i.e., not available in the real-world data experiments.

A. Evaluation on Synthetic Data

In this part, we design three groups of experiments. The first group is to verify whether the gap statistic is competent to selecting an appropriate tuning parameter s for ℓ_1- and $\ell_{1/2}$-c-means. The second group is to compare the performance of ℓ_1- and $\ell_{1/2}$-c-means with the related clustering approaches in the literature. The third one is to compare the approaches using data sets with a large number of groups. In the experiments, for FCM, AWFCM, and WEFCM, parameter α is selected from the fixed set $\alpha \in \{1.1, 1.2, \ldots, 3\}$ [8] following the procedure similar to [1], while parameter β is set to be 2 according to Keller and Klawonn's suggestion [8].

1) Simulation 1: We evaluate the performance of tuning parameter selection of ℓ_1- and $\ell_{1/2}$-c-means via the gap statistic [1], [27]. Assume the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has three clusters and each contains $n = 200$ objects and $p = 2000$ features. The first 50 features of $\mathbf{X}$ are relevant features while the others are noisy features. Assume $x_{ij} \sim \mathbb{N}(\mu_{ij}, 1)$ in $\mathbf{X}$ are independent and

$$\mu_{ij} = \begin{cases} \mu & \text{if } i \in C_1, j \leq 50 \\ 0 & \text{if } i \in C_2, j \leq 50 \\ -\mu & \text{if } i \in C_3, j \leq 50. \end{cases} \tag{18}$$

For $j > 50$, all noisy features follow the standard normal distribution $\mathbb{N}(0, 1)$. According to the setup, the relevant features have different mean values for different clusters, while the noisy features have the same distribution across all the clusters. We set $\mu = 0.2$, and repeat the simulation 100 times. Values of the tunning parameter s are selected to maximize the gap statistic.

Fig. 1 summarizes the results of ℓ_1- and $\ell_{1/2}$-c-means compared with the FCM separately. From the left subfigures, we can see that the highest gap statistic is achieved when the number of nonzero weights is around 50. It shows that the gap statistic is useful in selecting proper tuning parameter for ℓ_1- and $\ell_{1/2}$-c-means. The gap statistic of $\ell_{1/2}$-c-means is more sensitive to the selecting method comparing to that of ℓ_1-c-means, because the curve of gap statistic of $\ell_{1/2}$-c-means decreases dramatically when the highest gap

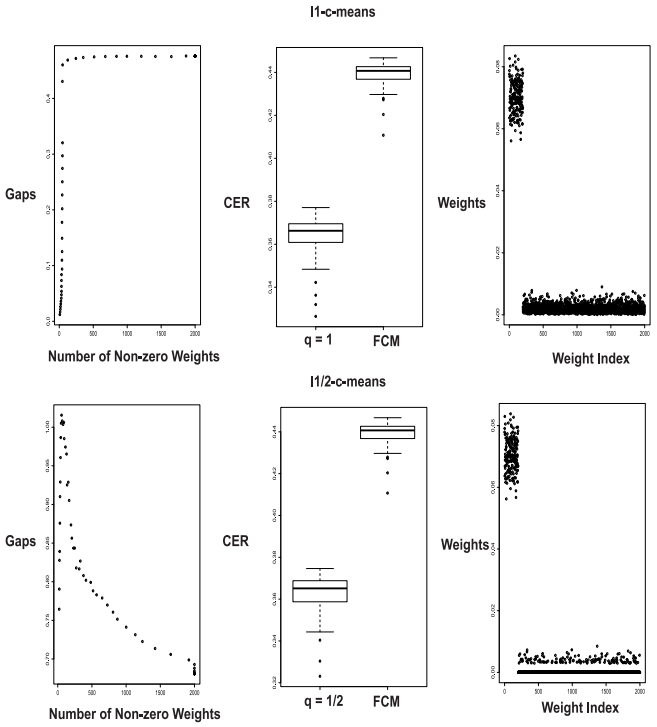


Fig. 1. ℓ_1-, $\ell_{1/2}$-c-means and FCM are applied to a simulated 3-class example. Left: the gap statistic versus the number of features with nonzero weights. Center: boxplots of the CERs. Right: weights obtained by the best s.

statistic is achieved. The middle subfigures show that the obtained partitions of ℓ_1- and $\ell_{1/2}$-c-means have significantly smaller CERs comparing to the standard FCM. The right subfigures report the average values of estimated feature weights over 100 trails. We can see that the average values for relevant features of ℓ_1- and $\ell_{1/2}$-c-means are all bigger than the noisy features. It shows that the use of gap statistic for ℓ_1- and $\ell_{1/2}$-c-means can help in selecting relevant features and improving the accuracy of partitions.

2) Simulation 2: In this experiment, we compare ℓ_1 and $\ell_{1/2}$-c-means with the aforementioned five related clustering approaches. Assume the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has six clusters and each cluster contains 200 objects and p features. The first 50 features of $\mathbf{X}$ are relevant features while the others are noisy features. Assume $x_{ij} \sim \mathbb{N}(\mu_{ij}, 1)$ in $\mathbf{X}$ are independent and

$$\mu_{ij} = \begin{cases} -\mu & \text{if } i \in C_1, j \le 50 \\ 0 & \text{if } i \in C_2, j \le 50 \\ \mu & \text{if } i \in C_3, j \le 50 \\ 2\mu & \text{if } i \in C_4, j \le 50 \\ 3\mu & \text{if } i \in C_5, j \le 50 \\ 4\mu & \text{if } i \in C_6, j \le 50. \end{cases} \tag{19}$$

For $j > 50$, all noisy features follow standard normal distribution $\mathbb{N}(0, 1)$. We set $\mu = 0.6, 0.8$, $p = 200, 500, 1000$ and repeat each simulation 100 times. The averaged clustering performance indices are shown in Table II.

Based on the average and standard deviation values in Table II, we conduct T-tests to examine whether the performance (in terms of CER, DB, and DI) are significantly different among the approaches, especially between our approaches (ℓ_1- and $\ell_{1/2}$-c-means) and the other five related clustering approaches. Our results indicate that CER and DB indices of FCM, WEFCM, AWFCM, and k-means are significantly higher and DI indices of FCM, WEFCM, AWFCM, and k-means are significantly lower comparing to ℓ_1- and $\ell_{1/2}$-c-means in most of the cases. Since FCM and k-means treat all features equally, and the solutions of WEFCM and AWFCM are linear combinations of all features (including noisy features), the noisy features negatively influence the clustering performance. Comparing to FCM, WEFCM, AWFCM, and k-means, the ℓ_1-k-means [1], ℓ_1- and $\ell_{1/2}$-c-means properly considered the noisy features, thus have relatively lower CER and DB values and higher DI values. Moreover, ℓ_1- and $\ell_{1/2}$-c-means approaches neither radically eliminate noisy features nor conservatively keep relevant features (see PZW and PNW), thus can produce even better clustering outputs than the ℓ_1-k-means.

3) Simulation 3: In this experimental study, we compare ℓ_1- and $\ell_{1/2}$-c-means with the five related clustering

TABLE II
MEAN AND STANDARD DEVIATION OF CER, DB INDEX, AND DUNN INDEX, MEAN VALUES OF PZW AND PNW FOR DIFFERENT MODELS IN EXPERIMENT 2

Approaches	CER	DB	DI	PZW	PNW
		$\mu = 0.6$, $p = 200$			
ℓ_1-c-means	**0.143(0.047)**	1.464(0.036)	0.143(0.031)	16	10
$\ell_{1/2}$-c-means	0.148(0.044)	**1.340(0.033)**	**0.194(0.054)**	126	50
FCM	0.333(0.057)	1.803(0.045)	0.132(0.032)	0	50
k-means	0.368(0.052)	1.992(0.021)	0.162(0.050)	0	50
ℓ_1-k-means	0.166(0.033)	1.537(0.045)	0.148(0.046)	150	27
WEFCM	0.276(0.048)	1.491(0.045)	0.188(0.047)	0	50
AWFCM	0.307(0.033)	1.504(0.037)	0.179(0.056)	0	50
		$\mu = 0.6$, $p = 500$			
ℓ_1-c-means	**0.123(0.034)**	1.258(0.035)	0.343(0.038)	345	50
$\ell_{1/2}$-c-means	0.232(0.033)	1.239(0.003)	**0.397(0.021)**	190	50
FCM	0.354(0.032)	1.597(0.031)	0.296(0.046)	0	50
k-means	0.389(0.055)	1.719(0.035)	0.203(0.052)	0	50
ℓ_1-k-means	0.302(0.045)	1.138(0.034)	0.245(0.047)	450	12
WEFCM	0.309(0.048)	**1.134(0.033)**	0.377(0.037)	0	50
AWFCM	0.302(0.045)	1.272(0.046)	0.366(0.057)	0	50
		$\mu = 0.6$, $p = 2000$			
ℓ_1-c-means	0.279(0.064)	1.424(0.059)	0.647(0.061)	892	50
$\ell_{1/2}$-c-means	**0.145(0.043)**	**1.301(0.057)**	0.674(0.037)	1950	36
FCM	0.378(0.043)	1.589(0.049)	0.588(0.033)	0	50
k-means	0.404(0.058)	1.553(0.043)	0.500(0.047)	0	50
ℓ_1-k-means	0.294(0.046)	1.332(0.072)	0.605(0.033)	1932	13
WEFCM	0.276(0.058)	1.437(0.037)	0.610(0.064)	0	50
AWFCM	0.370(0.045)	1.435(0.056)	**0.681(0.055)**	0	50
		$\mu = 0.8$, $p = 200$			
ℓ_1-c-means	0.263(0.034)	1.605(0.035)	**0.675(0.056)**	102	50
$\ell_{1/2}$-c-means	**0.213(0.045)**	**1.504(0.047)**	0.657(0.035)	145	49
FCM	0.278(0.049)	1.618(0.034)	0.667(0.047)	0	50
k-means	0.332(0.033)	1.606(0.039)	0.439(0.034)	0	50
ℓ_1-k-means	0.264(0.055)	1.586(0.038)	0.587(0.073)	150	16
WEFCM	0.312(0.049)	1.513(0.044)	0.672(0.037)	0	50
AWFCM	0.346(0.045)	1.537(0.034)	0.655(0.035)	0	50
		$\mu = 0.8$, $p = 500$			
ℓ_1-c-means	0.218(0.001)	1.461(0.002)	1.352(0.002)	437	10
$\ell_{1/2}$-c-means	**0.132(0.034)**	1.449(0.045)	1.461(0.058)	412	50
FCM	0.278(0.048)	1.451(0.033)	1.310(0.045)	0	50
k-means	0.392(0.034)	1.509(0.055)	1.287(0.044)	0	50
ℓ_1-k-means	0.253(0.045)	1.554(0.044)	1.359(0.037)	443	34
WEFCM	0.313(0.064)	1.498(0.033)	1.354(0.057)	0	50
AWFCM	0.345(0.046)	**1.438(0.046)**	**1.482(0.033)**	0	50
		$\mu = 0.8$, $p = 2000$			
ℓ_1-c-means	0.279(0.046)	1.246(0.047)	0.734(0.032)	1396	35
$\ell_{1/2}$-c-means	**0.114(0.047)**	**1.224(0.047)**	**0.840(0.034)**	1918	50
FCM	0.374(0.055)	1.253(0.033)	0.760(0.034)	0	50
k-means	0.393(0.048)	1.328(0.056)	0.730(0.032)	0	50
ℓ_1-k-means	0.321(0.055)	1.317(0.037)	0.740(0.046)	1144	50
WEFCM	0.263(0.047)	1.253(0.046)	0.797(0.048)	0	50
AWFCM	0.279(0.045)	1.247(0.048)	0.787(0.034)	0	50

TABLE III
MEAN AND STANDARD DEVIATION OF CER, DB INDEX, AND DUNN INDEX, MEAN VALUES OF PZW AND PNW FOR SYNTHETIC DATA WITH 20 CLUSTERS

Approaches	CER	DB Index	Dunn Index	PZW	PNW
ℓ_1-c-means	0.193(0.053)	**1.012(0.047)**	0.025(0.048)	145	20
$\ell_{1/2}$-c-means	**0.166(0.063)**	1.073(0.042)	**0.029(0.048)**	148	30
FCM	0.247(0.039)	1.424(0.048)	0.024(0.046)	0	50
k-means	0.344(0.037)	1.072(0.039)	0.027(0.037)	0	50
ℓ_1-k-means	0.190(0.043)	1.468(0.033)	0.026(0.045)	145	23
AWFCM	0.183(0.031)	1.067(0.037)	0.029(0.044)	0	50
WEFCM	0.220(0.038)	1.074(0.045)	0.027(0.045)	0	50

TABLE IV
SUMMARY OF UCI DATA SETS

Dataset	n	p	K
Yeast	1484	8	10
Libra Movement	360	90	15
Gesture Phase Segmentation	1747	50	7

TABLE V
EVALUATION RESULTS ON UCI DATA SETS

Dataset	Approaches	CER	DB	DI	NW
Yeast	$\ell_{1/2}$-c-means	0.264	0.510	**0.028**	3
	ℓ_1-c-means	0.249	0.376	0.024	5
	FCM	0.268	0.282	0.021	8
	kmeans	0.288	0.308	0.024	8
	ℓ_1-k-means	0.265	0.575	0.026	8
	AWFCM	**0.237**	0.269	**0.028**	8
	WEFCM	0.280	**0.266**	0.026	8
Libra Movement	$\ell_{1/2}$-c-means	0.098	**1.362**	0.097	15
	ℓ_1-c-means	0.106	1.373	**0.105**	19
	FCM	0.110	1.560	0.068	90
	k-means	0.091	1.477	0.051	90
	ℓ_1-k-means	**0.075**	1.479	0.069	28
	AWFCM	0.093	1.432	0.077	90
	WEFCM	0.088	1.462	0.051	90
Gesture Phase Segmentation	$\ell_{1/2}$-c-means	**0.352**	**0.479**	**0.032**	27
	ℓ_1-c-means	0.376	0.510	0.029	31
	FCM	0.377	0.549	0.030	50
	k-means	0.372	0.575	0.030	50
	ℓ_1-k-means	0.574	0.519	0.031	36
	AWFCM	0.359	0.508	**0.032**	50
	WEFCM	0.368	0.507	0.031	50

approaches by an example with a large number of clusters. Assume that the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ has $K = 20$ clusters and each cluster contains 200 objects and $p = 200$ features. Similar to simulation 2, we suppose the first 50 features of $\mathbf{X}$ are relevant features. We generate $x_{ij} \sim \mathbb{N}(\mu_{ij}, 1)$ in $\mathbf{X}$ independently. For the lth relevant features, we set $\mu_{ij} = (l-2)\mu, l = 1, \ldots, 50$. All the noisy features obey standard normal distribution $\mathbb{N}(0, 1)$. The experiment results are shown in Table III.

From Table III, we observe that $\ell_{1/2}$-c-means achieves the lowest CER and highest DI value. This indicates the outperformed clustering capability of $\ell_{1/2}$-c-means. Moreover, we also find that $\ell_{1/2}$-c-means approach neither radically eliminate noisy features nor conservatively keep relevant features (see PZW and PNW).

B. Evaluation on UCI Data Sets

In this section, three real-world UCI data sets [38] are used to evaluate the performance of the clustering approaches. The data sets are summarized in Table IV.

Table V summarizes the clustering results on the three UCI data sets. For the Yeast data set, we find that the AWFCM and WEFCM approaches rather than our ℓ_1- and $\ell_{1/2}$-c-means approaches have relatively better clustering performance in terms of the CER, DB, and DI values. This is because the Yeast data set is a low-dimensional data set with only eight features. The ℓ_1- and $\ell_{1/2}$-c-means selected only a part of the features as relevant features, thus may loss the information of the data set. Comparing to the Yeast data set, the Libra Movement and Gesture Phase Segmentation data sets all have relatively more features (90 and 50). On these two high-dimensional data sets, ℓ_1- and $\ell_{1/2}$-c-means perform relatively better (in terms of CER, DB, and DI), and identify the least number of relevant features. The identification of relevant features could improve the interpretability of the clustering results.

C. Evaluation on Human Activity Recognition Data

We also evaluate ℓ_1- and $\ell_{1/2}$-c-means and the five related clustering approaches in the task of human activity recognition (HAR) using smartphone data [39]. HAR tries to identify a person's specific activities given a set of observations of the person's actions and the surrounding environment [39], [40]. The HAR database uses smartphones to monitor 30 volunteers' activities of daily living [41]. The volunteers' age varies from 19 to 48. Each volunteer performed six daily activities (*Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing*, and *Laying*) with a smartphone (Samsung Galaxy S II) on the waist. The data set has 2947 observations, each with 561 features. These features describe the sensor signals (accelerometer and gyroscope), which were preprocessed by noisy filters and then sampled in fixed-width sliding windows of 2.56 s and 50% overlap. The target of this HAR task is to cluster different activities automatically by using the collected sensor signals.

We apply the clustering approaches on the data set and present the confusion matrix of the clustering results. A confusion matrix is a specific table layout which visualizes the performance of an algorithm. We use **recall** and **precision** to evaluate the clustering results. In pattern recognition and information retrieval, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also known as sensitivity) is the fraction of relevant instances that are retrieved.

Table VI shows the confusion matrices of ℓ_1- and $\ell_{1/2}$-c-means. From the tables, we observe that *Walking*, *Walking Upstairs*, and *Walking Downstairs* are similar to each other, thus are hard to be distinguished. *Sitting* and *Standing* are similar to each other. Comparing to the other confusion matrices, we find that ℓ_1- and $\ell_{1/2}$-c-means perform better than the other related approaches.

Furthermore, Table VII shows the CER, DB, DI values, and NWs for all the competing approaches. From Table VII, we find that $\ell_{1/2}$-c-means has the lowest CER value, the second lowest DB value, and the second highest DI value; at the same time, $\ell_{1/2}$-c-means identifies the least number of relevant features. This result indicates that $\ell_{1/2}$-c-means does

TABLE VI
CONFUSION MATRIX OF THE CLUSTERING RESULTS ON THE TEST DATA USING THE ℓ_1- AND $\ell_{1/2}$-C-MEANS MODEL. ROWS REPRESENT THE ACTUAL CLUSTERS AND COLUMNS THE ESTIMATED CLUSTERS

ℓ_1-c-means							
	Walking	Upstairs	Downstairs	Standing	Sitting	Laying	**Recall%**
Walking	209	146	141	0	0	0	42.137
Upstairs	284	138	49	0	0	0	29.299
Downstairs	62	106	252	0	0	0	60.000
Standing	5	0	0	218	309	0	40.977
Sitting	2	0	0	137	352	0	71.690
Laying	6	0	0	5	0	526	97.795
Precision%	36.796	35.385	57.014	60.556	38.124	100	
$\ell_{1/2}$-c-means							
	Walking	Upstairs	Downstairs	Standing	Sitting	Laying	**Recall%**
Walking	220	72	204	0	0	0	44.455
Upstairs	334	56	81	0	0	0	11.889
Downstairs	74	47	299	0	0	0	71.190
Standing	2	0	0	299	231	0	56.203
Sitting	2	0	0	201	288	0	58.656
Laying	4	0	0	4	0	529	98.510
Precision%	52.516	41.143	51.199	59.325	55.491	99.250	

TABLE VII
CER, DB INDEX, DUNN INDEX AND NW OF HAR DATA

Approaches	CER	DB	DI	NW
ℓ_1-c-means	0.1522	1.2152	**0.2999**	374
$\ell_{1/2}$-c-means	**0.1474**	1.0844	0.2842	**316**
AWFCM	0.2676	**1.0332**	0.2315	561
WEFCM	0.2716	1.1076	0.2276	561
FCM	0.2932	1.2858	0.1972	561
k-means	0.2516	1.3731	0.1655	561
ℓ_1-k-means	0.2201	1.2487	0.1713	500

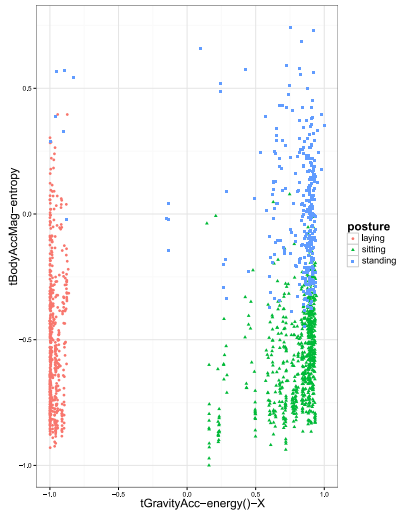


Fig. 2. Plotting relevant features [gGravityAcc-energy()-X and tBodyAccMag-entropy] for $\ell_{1/2}$-*c*-means.

not only correctly identify relevant features but also eliminate more noisy features. To further illustrate the performance of the $\ell_{1/2}$-*c*-means, here we show four features: two relevant features gGravityAcc-energy()-X and tBodyAccMag-entropy, which can be employed to distinguish the *Laying*, *Sitting*, and *Standing* clusters (as shown in Fig. 2), and two noisy features tBodyGyro-Correlation()-X,Z and angle(tBodyAccMean, gravity), which are irrelevant in this task (as shown in Fig. 3). $\ell_{1/2}$-*c*-means correctly identified the two relevant features

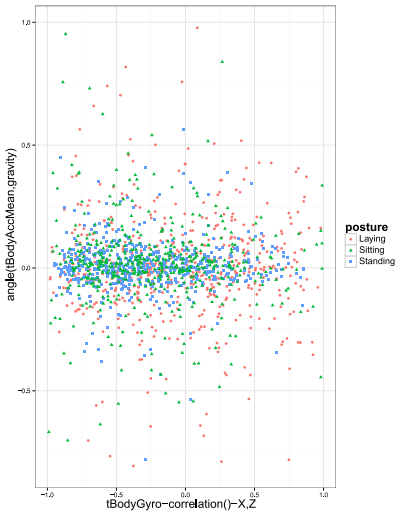


Fig. 3. Plotting noisy features [tBodyGyro-Correlation()-X,Z and angle(tBodyAccMean)] for $\ell_{1/2}$-*c*-means.

and eliminated the two noisy features. However, as to the competing approaches, the ℓ_1-*k*-means failed to identify the two relevant features; and ℓ_1-*k*-means, FCM, AWFCM, and WEFCM all treated the two noisy features as relevant features.

V. CONCLUSION

Clustering the high-dimensional data is challenging due to the existence of abundant noisy features. In this paper, inspired by the literature of sparse clustering, we proposed a novel FCM with sparse regularization (i.e., $\ell_q(0 < q \leq 1)$-norm regularization). To analytically concrete the model, ℓ_1- and $\ell_{1/2}$-*c*-means are proposed. We also developed an efficient algorithm to solve the model. The experimental results also confirmed the outperformance of our approach.

This paper has several limitations for further consideration. First, the proposed sparse FCM cannot be employed in cases, where clusters have different features. Generalizing the proposed sparse FCM to these cases is beyond the sparse clustering framework [1], which is still an open question. In future, our formulations could be extended to such scenarios. Second, the proposed sparse FCM is not applicable on big data applications. We will explore and discuss this problem based on a systematical framework proposed by Havens *et al.* [42]. Third, some other conventional and representative fuzzy clustering approaches (such as probabilistic *c*-means [5] and maximum entropy clustering [43]) could also be considered for handling high-dimensional data with the similar sparse regularization techniques.

APPENDIX
PROOFS

A. Proof of Theorem 1

If $\mathbf{w}$ and $\mathbf{C}$ are fixed, the model (7) can reduce to

$$\max_{\mathbf{C}} \; F(U) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}^{\alpha} \sum_{j=1}^{p} w_j \left(x_{ij} - c_{kj}\right)^2$$

$$\text{s.t.} \quad \sum_{k=1}^{K} u_{ik} = 1, 0 \le u_{ik} \le 1$$
$$i = 1, \ldots, n, k = 1, \ldots, K.$$

The cases $\sum_{j=1}^{p} w_j (x_{ij} - c_{kj})^2 = 0$ or $\sum_{j=1}^{p} w_j (x_{ij} - c_{kj})^2 \neq 0$ are trival. So we only consider the last case. The Lagrange function of $F(U)$ is

$$L(u_i, \lambda) = \sum_{i=1}^{n} \sum_{j=1}^{p} u_{ik}^{\alpha} w_j \left(x_{ij} - c_{kj}\right)^2 - \lambda \left(\sum_{k=1}^{K} u_{ik} - 1 \right)$$

where $u_i = (u_{i1}, \ldots, u_{iK})^{\top}$ and $i = 1, \ldots, n$. We take the partial derivative with respect to u_{ik} and λ, and set

$$\frac{\partial L(u_i, \lambda)}{\partial u_{ik}} = 0 \tag{20}$$

and

$$\frac{\partial L(u_i, \lambda)}{\partial \lambda} = 0. \tag{21}$$

From (20) we have

$$u_{ik} = \left[\frac{\lambda}{\alpha \sum_{j=1}^{p} w_j \left(x_{ij} - c_{kj}\right)^2} \right]^{\frac{1}{\alpha - 1}}. \tag{22}$$

Then substituting (22) to (21), we have

$$\left(\frac{\lambda}{\alpha} \right)^{\frac{1}{\alpha - 1}} = \frac{1}{\sum_{t=1}^{K} \left[\frac{1}{\sum_{j=1}^{p} w_j \left(x_{ij} - c_{tj}\right)^2} \right]^{\frac{1}{\alpha - 1}}}. \tag{23}$$

Finally, substituting (23) into (22) and we get

$$u_{ik} = \frac{1}{\sum_{t=1}^{K} \left[\frac{\sum_{j=1}^{p} w_j \left(x_{ij} - c_{kj}\right)^2}{\sum_{j=1}^{p} w_j \left(x_{ij} - c_{tj}\right)^2} \right]^{\frac{1}{\alpha - 1}}}. \tag{24}$$

B. Proof of Theorem 2

Let U and the attribute weights $\mathbf{w}$ be fixed, $F(U, \mathbf{C}, \mathbf{w})$ is reduced to

$$\max_{\mathbf{C}} F(\mathbf{C}) = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik}^{\alpha} \sum_{j=1}^{p} w_j \left(x_{ij} - c_{kj}\right)^2.$$

First, it is obvious that if $w_j = 0$, we can get $c_{kj} = 0$. If $w_j \neq 0$, we fix $\mathbf{w}$ and U, and the first order derivative of $F(\mathbf{C})$ is

$$\frac{\partial F(\mathbf{C})}{\partial c_{kj}} = 2 \sum_{i=1}^{n} u_{ik}^{\alpha} w_j \left(x_{ij} - c_{kj}\right) \tag{25}$$

and let $(\partial F(\mathbf{C}) / \partial c_{kj}) = 0$, we get $c_{kj} = (\sum_{i=1}^{n} u_{ik}^{\alpha} x_{ij} / \sum_{i=1}^{n} u_{ik}^{\alpha})$.

C. Proof of Theorem 4

Assume $\lambda_j^{*1/q} = w_j^*$, $j = 1, 2, \ldots, p$, is the optimal solution of the optimization problem. So we can derive that $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*, \ldots, \lambda_p^*)^{\intercal}$ is the local maximum solution of

$$\max_{\boldsymbol{\lambda}} \sum_{j=1}^{p} \lambda_j^{1/q} a_j$$
$$\text{s.t.} \sum_{j=1}^{p} \lambda_j^{2/q} \le 1, \sum_{j=1}^{p} \lambda_j \le s.$$

The Lagrange equation is

$$L = a_j \lambda_j^{1/q} + \mu_1 \left(\lambda_j^{2/q} - 1 \right) + \mu_2 \left(\lambda_j - s \right). \tag{26}$$

Thought the KKT Conditions, there exists $\mu_1 \le 0$ and $\mu_2 \le 0$ such that for any nonzero element u_j^*, and set the first derivative of L equal to 0

$$f_j(x_j) = \frac{1}{q} a_j x_j^{1/q - 1} + \frac{2}{q} \mu_1 x_j^{2/q - 1} + \mu_2 = 0. \tag{27}$$

Then we derive

$$f_j'(x) = \frac{1}{q}(1/q - 1) a_j x_j^{\frac{1}{q} - 2} + \frac{2}{q}(2/q - 1) \mu_1 x_j^{\frac{2}{q} - 2} \ge 0. \tag{28}$$

Thus, λ^* is the local minimal solution of problem (2) instead of maximum point, so μ_1 should be negative.

Since λ_j^* is the positive root of $f_j(x) = 0$, leting $f_j'(x) = 0$, we get

$$x_0 = \left[\frac{\left(-\frac{1}{q}(\frac{1}{q} - 1) a_j \right)}{\frac{2}{q}(\frac{2}{q} - 1) \mu_1} \right]^q = \left[\frac{(1 - q) a_j}{2(2 - q) \mu_1} \right]^q. \tag{29}$$

Because $f_j'(x) > 0$ when $x < x_0$, and $f_j'(x) < 0$ when $x > x_0$, so $f_j(x)$ has two non-negative real roots, noted as φ_{j1} and φ_{j2}. So $\lambda_j^* = \varphi_{j1}, \varphi_{j2}$ or 0. From $f_j'(x) = 0$ we can get

$$a_j^{2-q} > (2 - q) q \left(\frac{2(2 - q)}{1 - q} \right)^{1 - q} (-\mu_1)^{1 - q} (-\mu_2). \tag{30}$$

Define

$$g(x) \coloneqq q r^{\frac{1}{q} - 1} f\left(\frac{x}{r} \right) = a_j x^{\frac{1}{q} - 1} - 2x^{\frac{2}{q} - 1} - \frac{1}{(2 - q)} \left(\frac{1 - q}{2(2 - q)} \right)^{1 - q} \phi$$

where $\phi \coloneqq (2 - q) q ([2(2 - q)/(1 - q)])^{1 - q} (-\mu_1)^{1 - q} (-\mu_2)$ and $r = a_j^{2-q}$.

Since $r > 0$, we have that $r\varphi_{\gamma j}$ is the γth largest real root of $g(x)$, $\gamma = 1, 2$. Denote $F_{\gamma j}(\phi)$ is the real root of $a_j x^{(1/q) - 1} - 2x^{(2/q) - 1} - (1/(2 - q))([(1 - q)/2(2 - q)])^{1 - q} \phi = 0$, then $r\varphi_{\gamma j} = F_{\gamma j}(\phi)$, equivalent to $\varphi_{\gamma j} = (F_{\gamma j}(\phi)/r)$. So we have $\varphi_{1 j_1} \le \varphi_{1 j_2}, \forall j_1 < j_2$, and $\varphi_{2 j_1} \le \varphi_{2 j_2}, \forall j_1 < j_2$. Since λ^* is the optimal solution and $a_{j_1} \ge a_{j_2}, \forall j_1 < j_2$, we can derive that $\lambda_{j_1}^* \ge \lambda_{j_2}^*$. Denote k as the smallest index such that $\lambda_{k+1}^* = 0$. So we get $\lambda_j^* = \varphi_{1j}, \forall j \le k - 1$ and $\lambda_j^* = \varphi_{1k}$ or φ_{2k} and $\lambda_j^* = 0, \forall j > k$.

Here, we introduce an extra symbol Δ to simplify the formulas. If $\lambda_j^* = \varphi_{1k}$, let $\Delta = \sum_{j=k+1}^{p+1} 2a_j^{2-q} + \phi$; if $\lambda_j^* = \varphi_{2k}$, let $\Delta = \sum_{j=k}^{p+1} 2a_j^{2-q} - \phi$. Define

$$\phi(\Delta) = \begin{cases} \Delta - T \\ \text{if } T \le \Delta \le T + a_j^{2-q} \\ T - \Delta \\ \text{if } T + a_j^{2-q} \le \Delta \le T \end{cases}$$

where $h = p, p-1, \ldots, 1.$ and $T = 2\sum_{h=j+1}^{p+1} a_h^{2-q}$. Then

$$w_j^* = \begin{cases} F_{1j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 0 \le \Delta < 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \\ F_{2j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \le \Delta < 2\sum_{l=j}^{p+1} a_l^{2-q} \\ 0 \quad \text{if } \Delta \ge 2\sum_{l=j}^{p+1} a_l^{2-q} \end{cases}$$

where $j = 1, \ldots, p$ and r is the scaling factor to ensure $\|\mathbf{w}^*\|_2^2 = 1$.

D. Proof of Corollary 1

Based on Theorem 3, solving (10) for $q = 1$ is equivalent to find the optimal solution of following optimization problem:

$$\begin{aligned} \min_{\mathbf{w}} \ & -\mathbf{a}^\top \mathbf{w} \\ \text{s.t.} \ & \|\mathbf{w}\|_2^2 = 1, \|\mathbf{w}\|_1 = s \\ & w_j \ge 0, j = 1, \ldots, p. \end{aligned}$$

Consider the KKT condition of the above optimization problem

$$-\mathbf{a} + 2\lambda_1 \mathbf{w} + \lambda_2 \vec{1} - v = 0 \tag{31}$$

$$\|\mathbf{w}\|_2^2 = 1 \tag{32}$$

$$\|\mathbf{w}\|_1 = s \tag{33}$$

$$\mathbf{w} \succeq 0 \tag{34}$$

$$\lambda_1, \lambda_2 > 0, v \succeq 0 \tag{35}$$

$$v_j w_j = 0 \tag{36}$$

where $\mathbf{w} \succeq 0$ means $w_j \ge 0, \forall j = 1, \ldots, p$. From (31), we have $\mathbf{a} - \lambda_2 \vec{1} = 2\lambda_1 \mathbf{w} - v$. Since we know (36), then $v_j = 0$ if for some $w_j > 0$. Furthermore, $2\lambda_1 w_j = a_j - \lambda_2$. Using the same trick, $v_j = \lambda_2 - a_j$ if for some $v_j > 0$, that is $w_j = 0$. Then, $\mathbf{w} = (1/2\lambda_1) S(a, \Delta)$ where we choose $\Delta := \lambda_2$. Because $\|\mathbf{w}\|_2 = 1$, then $\mathbf{w}^* = (S(a, \Delta)/\|S(a, \Delta)\|_2)$.

E. Proof of Corollary 2

According to Theorem 4, for any $0 < q < 1$, we have the optimal solution of (10) satisfies

$$w_j^* = \begin{cases} F_{1j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 0 \le \Delta < 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \\ F_{2j}(\varphi(\Delta))/r^{1/p} \\ \quad \text{if } 2\sum_{l=j+1}^{p} a_l^{2-q} + a_j^{2-q} \le \Delta < 2\sum_{l=j}^{p+1} a_l^{2-q} \\ 0 \quad \text{if } \Delta \ge 2\sum_{l=j}^{p+1} a_l^{2-q} \end{cases}$$

where $j = 1, \ldots, p$, r is the scaling factor to ensure $\|\mathbf{w}^*\|_2 = 1$, and $F_{\gamma j}(\phi)$ is the γth largest real root of equation

$$a_j x_j^{1-q} - 2x_j^{2-q} - \left(\frac{1-q}{2(2-q)}\right)^{1-q} \frac{1}{2-q}\phi = 0. \tag{37}$$

For $q = 1/2$, (37) becomes

$$2u_j^3 - a_j u_j + \frac{2}{3\sqrt{6}}\phi = 0$$

where $u_j = x_j^{1/2} \ge 0$. Then, based on the famous Cartan formula [19], we can obtain the two non-negative solutions of the cubic equation as

$$F_{\gamma j}(\phi) = \frac{2}{3} a_j \cos^2\left(\frac{\pi}{3} + (-1)^\gamma \frac{1}{3} \arccos \frac{\phi}{a_j^{3/2}}\right)$$

where $\gamma = 1, 2$.

REFERENCES

[1] 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.

[2] Z. Deng, K.-S. Choi, Y. Jiang, J. Wang, and S. Wang, "A survey on soft subspace clustering," *Inf. Sci.*, vol. 348, pp. 84–106, Jun. 2016.

[3] W. Pan and X. Shen, "Penalized model-based clustering with application to variable selection," *J. Mach. Learn. Res.*, vol. 8, pp. 1145–1164, May 2007.

[4] X. Chang, Y. Wang, R. Li, and Z. Xu, "Sparse k-means with ℓ_0/ℓ_∞ penalty for high-dimensional data clustering," *arXiv preprint arXiv:1403.7890*, 2014.

[5] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[6] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 517–530, Aug. 2005.

[7] L. Zhu, F.-L. Chung, and S. Wang, "Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 3, pp. 578–591, Jun. 2009.

[8] A. Keller and F. Klawonn, "Fuzzy clustering with weighting of data variables," *Int. J. Uncertainty Fuzziness Knowl. Based Syst.*, vol. 8, no. 6, pp. 735–746, 2000.

[9] C. Tang, S. Wang, and W. Xu, "New fuzzy c-means clustering model based on the data weighted approach," *Data Knowl. Eng.*, vol. 69, no. 9, pp. 881–900, 2010.

[10] E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang, "An optimization algorithm for clustering using weighted dissimilarity measures," *Pattern Recognit.*, vol. 37, no. 5, pp. 943–952, 2004.

[11] Y. Jiang *et al.*, "Collaborative fuzzy clustering from multiple weighted views," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 688–701, Apr. 2015.

[12] P. Qian *et al.*, "Cluster prototypes and fuzzy memberships jointly leveraged cross-domain maximum entropy clustering," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 181–193, Jan. 2016.

[13] L. Peng and J. Zhang, "An entropy weighting mixture model for subspace clustering of high-dimensional data," *Pattern Recognit. Lett.*, vol. 32, no. 8, pp. 1154–1161, 2011.

[14] Z. Deng *et al.*, "Transfer prototype-based fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 24, no. 5, pp. 1210–1232, Oct. 2016.

[15] J. Zhou and C. L. P. Chen, "Attribute weight entropy regularization in fuzzy c-means algorithm for feature selection," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Macau, China, 2011, pp. 59–64.

[16] W.-C. Chang, "On using principal components before separating a mixture of two multivariate normal distributions," *Appl. Stat.*, vol. 32, no. 3, pp. 267–275, 1983.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] Y. Liu, H. H. Zhang, C. Park, and J. Ahn, "Support vector machines with adaptive Lq penalty," *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 6380–6394, 2007.

[19] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$l_{1/2}$ regularization: A thresholding representation theory and a fast solver," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1013–1027, Jul. 2012.

[20] J. Zeng, Z. Xu, B. Zhang, W. Hong, and Y. Wu, "Accelerated $l_{1/2}$ regularization based SAR imaging via BCR and reduced Newton skills," *Signal Process.*, vol. 93, no. 7, pp. 1831–1844, 2013.

[21] X. Chang, Z. Xu, H. Zhang, J. Wang, and Y. Liang, "Robust regularization theory based on L_q (0<q<1) regularization: The asymptotic distribution and variable selection consistence of solutions," *Scientia Sinica Mathematica*, vol. 40, no. 10, pp. 985–998, 2010.

[22] G. Marjanovic and V. Solo, "On l_q optimization and matrix completion," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5714–5724, Nov. 2012.

[23] L. Niu, R. Zhou, Y. Tian, Z. Qi, and P. Zhang, "Nonsmooth penalized clustering via ℓ_p regularized sparse regression," *IEEE Trans. Cybern.*, to be published. [Online]. Available: http://ieeexplore.ieee.org/document/7460120/

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B Methodol.*, vol. 58, no. 1, pp. 267–288, 1996.

[25] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[26] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[27] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *J. Roy. Stat. Soc. B Stat. Methodol.*, vol. 63, no. 2, pp. 411–423, 2001.

[28] G. Gan and J. Wu, "A convergence theorem for the fuzzy subspace clustering (FSC) algorithm," *Pattern Recognit.*, vol. 41, no. 6, pp. 1939–1947, 2008.

[29] J. H. Friedman and J. J. Meulman, "Clustering objects on subsets of attributes (with discussion)," *J. Roy. Stat. Soc. B Stat. Methodol.*, vol. 66, no. 4, pp. 815–849, 2004.

[30] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in K-means type clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 657–668, May 2005.

[31] L. Jing, M. K. Ng, and J. Z. Huang, "An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 8, pp. 1026–1041, Aug. 2007.

[32] C.-Y. Tsai and C.-C. Chiu, "Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm," *Comput. Stat. Data Anal.*, vol. 50, no. 10, pp. 4658–4672, 2008.

[33] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, 1957.

[34] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.*, vol. 3, no. 3, pp. 32–57, 1973.

[35] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Probab.*, vol. 1. Berkeley, CA, USA, 1967, pp. 281–297.

[36] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979.

[37] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, no. 1, pp. 95–104, 1974.

[38] K. Bache and M. Lichman. (2013). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml

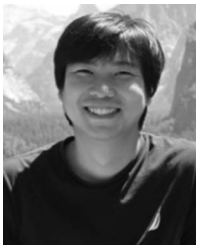
[39] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn. (ESANN)*, Bruges, Belgium, 2013, pp. 437–442.

[40] P. Lukowicz *et al.*, "Recognizing workshop activity using body worn microphones and accelerometers," in *Pervasive Computing*. Heidelberg, Germany: Springer, 2004, pp. 18–32.

[41] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Energy efficient smartphone-based activity recognition using fixed-point arithmetic," *J. Univ. Comput. Sci.*, vol. 19, no. 9, pp. 1295–1314, 2013.

[42] T. C. Havens, J. C. Bezedk, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy *c*-means algorithms for very large data," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 6, pp. 1130–1146, Dec. 2012.

[43] N. B. Karayiannis, "MECA: Maximum entropy clustering algorithm," in *Proc. 3rd IEEE Conf. Fuzzy Syst. IEEE World Congr. Comput. Intell.*, Orlando, FL, USA, 1994, pp. 630–635.



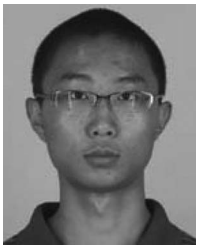
Xiangyu Chang received the Ph.D. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2014.

He is currently an Assistant Professor with the School of Management, Xi'an Jiaotong University. His current research interests include statistical machine learning, social network analysis, high-dimensional statistics, and business statistics.

Qingnan Wang received the B.Sc. degree in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2013, where she is currently pursuing the Ph.D. degree with the Department of Information System and E-Business, School of Management.

Yuewen Liu received the Ph.D. degree in information systems from the City University on Hong Kong, Hong Kong, in 2009, and the Ph.D. degree in management science and engineering from the University of Science and Technology of China, Hefei, China, in 2010.

He is currently an Assistant Professor with the School of Management, Xi'an Jiaotong University, Xi'an, China. His current research interests include social network analysis and e-commerce and business statistics.



Yu Wang received the B.Sc. and M.A. degrees in applied mathematics from Xi'an Jiaotong University, Xi'an, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Department of Statistics, University of California at Berkeley, Berkeley, CA, USA.

From 2007 to 2009, he was a member of the Special Class of the Gifted Young in Xi'an Jiaotong University. From 2009 to 2013, he was in the Science Topnotch Program in China.