Applied Mathematics and Computation 382 (2020) 125328

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Sparse probabilistic K-means

Yoon Mo Jung[a], Joyce Jiyoung Whang[b], Sangwoon Yun[c,*]

[a] Department of Mathematics, Sungkyunkwan University, Suwon 16419, Republic of Korea
[b] Department of Computer Science and Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
[c] Department of Mathematics Education, Sungkyunkwan University, Seoul 03063, Republic of Korea

ARTICLE INFO

Article history:
Received 4 October 2019
Revised 20 March 2020
Accepted 19 April 2020

Keywords:
Clustering
K-means
Alternating minimization

ABSTRACT

The goal of clustering is to partition a set of data points into groups of similar data points, called clusters. Clustering algorithms can be classified into two categories: hard and soft clustering. Hard clustering assigns each data point to one cluster exclusively. On the other hand, soft clustering allows probabilistic assignments to clusters. In this paper, we propose a new model which combines the benefits of these two models: clarity of hard clustering and probabilistic assignments of soft clustering. Since the majority of data usually have a clear association, only a few points may require a probabilistic interpretation. Thus, we apply the ℓ_1 norm constraint to impose sparsity on probabilistic assignments. Moreover, we also incorporate outlier detection in our clustering model to simultaneously detect outliers which can cause serious problems in statistical analyses. To optimize the model, we introduce an alternating minimization method and prove its convergence. Numerical experiments and comparisons with existing models show the soundness and effectiveness of the proposed model.

1. Introduction

Clustering is one of the most fundamental problems in unsupervised learning and is often the basis for further learning. The goal of clustering is to partition a set of data points into groups, called clusters, so that similar data points are assigned to the same cluster. The clustering problem has been addressed in various contexts and disciplines, such as machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, and computer graphics, just to name a few.

Due to its long history and diverse applications, there are tremendous algorithms and works in the literature and they can be classified as several ways: hard and soft clustering, hierarchical and partitional clustering, by distance and similarity measure, using pdf or mixture densities, graph-based, neural networks-based, kernel-based, and so on. For a comprehensive survey and classification of clustering algorithms, we refer the review paper [1].

The most popular method and the basic prototype is k-means; many other algorithms are motivated from it or related to it. The k-means algorithm is a hard clustering algorithm which assigns each data point to one cluster exclusively. However, in real datasets, this assignment might be inappropriate when grouping is obscure or overlapped in the data. To allow overlaps between clusters, soft clustering is often applied. The idea of soft clustering is to assign a probability of each data point to a cluster so that each object may belong to multiple clusters with a certain degree. Fuzzy c-means is a well-knwon

[*] Corresponding author.
E-mail addresses: ymjung@skku.edu (Y.M. Jung), jjwhang@skku.edu (J.J. Whang), yswmathedu@skku.edu (S. Yun).

https://doi.org/10.1016/j.amc.2020.125328
0096-3003/© 2020 Elsevier Inc. All rights reserved.

soft clustering algorithm, which applies concepts in fuzzy logic and fuzzy set theory. The most widely used fuzzy clustering algorithm can be founded in [2]. In this case, most data points are associated with multiple clusters, even when their clusters are feasible and it may introduce extra fuzziness in interpretation.

In this paper, we combine the benefits of the two different clustering approaches: clarity of hard clustering and probabilistic assignments of soft clustering. For example, let us consider assigning a number between 0 and 1 to indicate the clusters 0 and 1. When the affiliation of a data point is clear, it is desirable to assign the discrete values 0 and 1. Only when that is not clear, it is proper to assign some neutral values such as 0.5. If such affiliation is expressed by a matrix and the discrete values 0 and 1 are neglected, the matrix is sparse. To promote exclusive assignments on the majority of data points which have clear associations, we add the ℓ_1 norm in our clustering model. Owing to sparsity in solutions, the ℓ_1 norm constraint is widely used in compressed sensing, image processing, statistics and machine learning [3–5]. With the ℓ_2 norm in prevalent clustering models, we intend to enhance the prediction and interpretability of the models. Actually, coupling penalty terms are popular and successful in many branches of applied mathematics [6–8].

In addition, datasets are often contaminated by outliers which should not belong to any cluster. Outliers are caused by high variability in data, measurement errors, faulty data, erroneous procedures, for example. Since they can cause serious problems in statistical analyses, outlier detection has an extensive literature. Even though outliers are defined by a set of data points that do not belong to any *cluster*, most existing clustering algorithms do not incorporate the outlier detection task in the clustering model. Recently, there have been several efforts to handle clustering and outlier detection simultaneously [9–11]. Along with this direction of research, we also propose a model which simultaneously incorporates probabilistic assignments (or overlapping clustering) and outlier detection.

This paper is organized as follows. Section 2 formulates clustering problems in standpoint of k-means and presents a few models related to this paper. Section 3 proposes new models which allow overlapping on only a few points. In addition, we also combine clustering and outlier detection in a unified way. We also show the properties of our model by investigating eigenvectors. The optimization procedures and convergence are given in Section 4 and numerical results on synthetic and real data are given in Section 5. Finally, Section 6 concludes the paper.

2. Related models

Given a set of n data points $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_n\}$ with $\mathbf{x}_j \in \mathbb{R}^d$, k-means [12] seeks a partition of the data points into k clusters $\mathbf{C} = \{C_1, \ldots, C_k\}$ such that they cover all points, i.e., $\cup_{i=1}^k C_i = \mathbf{X}$ and partitions are mutually disjoint, i.e., $C_i \cap C_{i'} = \emptyset$ for $i \neq i'$. For that purpose, the k-means algorithm minimizes the sum of squared distances from data points to their cluster centroids. The objective of k-means is defined as follows:

$$\min_{\mathbf{C},\mathbf{m}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2,$$

where $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_k)$. Here, $\|\cdot\|$ denotes the Euclidian norm otherwise specified. Considering only minimizing $\sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2$, $\mathbf{m}_i$ is the centroid of the cluster C_i. So, the problem is equivalent to

$$\min_{\mathbf{C}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \mathbf{m}_i\|^2 \quad \text{where} \quad \mathbf{m}_i = \frac{\sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j}{|C_i|}. \tag{1}$$

Finding a global minimum of (1) is an NP-hard problem, due to partitioning. Thus, only a local optimum is sought by computational algorithms. A well-known heuristic algorithm, called Lloyd's algorithm [12], which alternates finding centroids of clusters and assigning data points to their closest clusters, is commonly employed.

To represent the partitions, often a membership matrix $U = [u_{ij}]_{k\times n}$ is introduced:

$$u_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_j \in C_i, \\ 0, & \text{otherwise.} \end{cases}$$

Using the membership matrix U, The k-means clustering can be rewritten as:

$$\min_{U} \sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} \|\mathbf{x}_j - \mathbf{m}_i\|^2 \quad \text{where} \quad \mathbf{m}_i = \frac{\sum_{j=1}^n u_{ij}\mathbf{x}_j}{\sum_{j=1}^n u_{ij}}.$$

Thus, the k-means clustering problem is also regarded as finding a membership matrix $U = [u_{ij}]_{k\times n}$ with the properties:

$$u_{ij} = 0 \text{ or } 1, \quad \text{and} \quad \sum_{i=1}^{k} u_{ij} = 1.$$

One another feasible constraint is $\sum_{j=1}^n u_{ij} < n$, which excludes the case of assigning all the data points into one cluster; we assume this constraint in the below without an explicit statement.

The constraint $u_{ij} = 0$ or 1 is often relaxed as $0 \le u_{ij} \le 1$; the former is called *hard clustering* and the latter is *soft clustering* where u_{ij} can be interpreted as a probability that a data point j belongs to cluster i. Fuzzy c-means (or fuzzy k-means) [2] is a well-known soft clustering method, and the optimization problem is defined as:

$$\begin{array}{ll} \min_{U,\mathbf{m}} & \sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij})^r\|\mathbf{x}_j - \mathbf{m}_i\|^2 \\ & \text{subject to} \quad u_{ij} \ge 0, \quad 1 \le i \le c,\ 1 \le j \le n, \\ & \qquad\qquad \sum_{i=1}^{c} u_{ij} = 1, \quad 1 \le i \le c, \end{array}$$

with exponent $r \ge 1$. In this model, $\mathbf{m}_i$ may be evaluated as $\frac{\sum_{j=1}^{n}(u_{ij})^r\mathbf{x}_j}{\sum_{j=1}^{n}(u_{ij})^r}$. Thus, Lloyd type alternating iteration algorithms can be applicable. A standard choice of r is 2 since it is closely related to popular methods such as least squares and ℓ_2 norm optimization.

In real-world datasets, the clusters can be overlapped with each other, and there exist outliers that do not belong to any cluster. The NEO-K-Means method has been proposed to deal with these real-world clustering problems [9]. The NEO-K-Means model assumes a hard clustering, i.e., $u_{ij} = 0$ or 1. Each column of U may have multiple ones (to allow overlaps between clusters) or all zeros (to allow outliers). With the vector of all ones $\mathbf{1}$ and the indicator function $\mathbb{I}\{\cdot\}$, the NEO-K-Means clustering is defined as follows:

$$\begin{array}{ll} \min_U & \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\|\mathbf{x}_j - \mathbf{m}_i\|^2 \quad \text{where} \quad \mathbf{m}_i = \frac{\sum_{j=1}^{n} u_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}}, \\ & \text{subject to} \quad \text{trace}(UU^T) = (1+\alpha)n,\ \sum_{j=1}^{n}\mathbb{I}\{(\mathbf{1}^T U)_j = 0\} \le \beta n, \end{array} \tag{2}$$

where $0 \le \alpha \ll k-1$ and $0 \le \beta \ll n$. The parameter α determines the amount of overlap and β controls the number of outliers (note that the second constraint counts the number of zero columns). If $\alpha = \beta = 0$, (2) becomes equivalent to the k-means clustering.

3. Proposed models

The standard fuzzy c-means, i.e., that with exponent $r = 2$ tends to produce a dense matrix U, which is a typical property of ℓ_2 norm-based optimizations. On the other hand, ℓ_1 norm-based optimizations often induce sparsity in solutions, and thus are widely used in compressed sensing, image processing, statistics and machine learning communities [3–5]. For example, lasso (least absolute shrinkage and selection operator) [5] applies ℓ_1 norm for variable selection and regularization to enhance the prediction accuracy and interpretability of statistical models. Comparing other popular shrinkage methods, ridge regression which uses ℓ_2 norm, for example, lasso returns only a few nonzero variables [6].

To allow multi-cluster assignment but to promote one or few memberships, one may consider fuzzy c-means with exponent $r = 1$:

$$\begin{array}{ll} \min_{U,\mathbf{m}} & \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\|\mathbf{x}_j - \mathbf{m}_i\|^2 \\ & \text{subject to} \quad u_{ij} \ge 0, \quad 1 \le i \le k,\ 1 \le j \le n, \\ & \qquad\qquad \sum_{i=1}^{k} u_{ij} = 1, \quad 1 \le j \le n. \end{array} \tag{3}$$

The problem is equivalent to minimizing the objective with respect to U by setting $\mathbf{m}_i = \frac{\sum_{j=1}^{n} u_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}}$ [2]. However, the formulation (3) is not really probabilistic. The condition $u_{ij} = 0$ or 1 is relaxed to $u_{ij} \in [0, 1]$, but u_{ij} must be exclusively 0 or 1, if the distance squares to centroids $\|\mathbf{x}_j - \mathbf{m}_i\|^2$ are different. That is, the solution of (3) is somewhat equivalent to that of k-means [2]. Lemma 3.1 shows this property.

Lemma 3.1. *We fix the values of $\mathbf{m}_i$, $1 \le i \le k$ or assume that they have the values at a local minimum. If $\|\mathbf{x}_j - \mathbf{m}_i\|^2$ are not same for all i, then $u_{ij} = 0$ or 1.*

Proof. Since the values of $\mathbf{m}_i$ are fixed, the objective $\sum_{j=1}^{n}\sum_{i=1}^{k} u_{ij}\|\mathbf{x}_j - \mathbf{m}_i\|^2$ is separable with respect to j, we only consider $\sum_{i=1}^{k} u_{ij}\|\mathbf{x}_j - \mathbf{m}_i\|^2$ for a fixed j. By replacing u_{ij} and $\|\mathbf{x}_j - \mathbf{m}_i\|^2$ by x_i and c_i, respectively, we have the following linear programming:

$$\begin{array}{l} \min c_1x_1 + \cdots + c_kx_k \\ \text{subject to} \quad x_i + \cdots + x_k = 1, \quad x_i \ge 0,\ 1 \le i \le k. \end{array}$$

If $y = c_1x_1 + \cdots + c_kx_k$ is not parallel to $x_i + \cdots + x_k = 1$, i.e., $c_1 = \ldots = c_k$ does not hold, the minimum occurs at a vertex $\mathbf{e}_l = (0, \ldots, 1, \ldots, 0)$ for some $l = 1, \ldots, k$ of the standard simplex by the constraint. □

To promote probabilistic assignments (i.e., soft clustering), we add an ℓ_2 penalty term, similar to the elastic net [13] which linearly combines the ℓ_1 and ℓ_2 penalties of the lasso and ridge regression: With $\lambda \ge 0$,

$$\begin{array}{ll} \min_{U,\mathbf{m}} & \sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}\|\mathbf{x}_j - \mathbf{m}_i\|^2 + \lambda\sum_{i=1}^{k}\sum_{j=1}^{n} u_{ij}^2 \\ & \text{subject to} \quad u_{ij} \ge 0, \quad 1 \le i \le k,\ 1 \le j \le n, \\ & \qquad\qquad \sum_{i=1}^{k} u_{ij} = 1, \quad 1 \le j \le n. \end{array} \tag{4}$$

We call the model (4) Sparse Probabilistic k-means (SP K-MEANS). Notice that the minimizer of $u_1^2+u_2^2$ subject to $u_1+u_2=1$ and $u_1, u_2 \geq 0$ is $u_1=u_2=\frac{1}{2}$. Thus, if the k-means term in the objective is neglected or has no significant role, an indifferent value $\frac{1}{2}$ is assigned for $k=2$. For general k, $\frac{1}{k}$ will be selected. The parameter λ controls the balance between exclusive assignment by the k-means term and uncertainty by the second term and the model approaches to the standard k-means as λ decreases to 0. The next theorem shows the relation between λ and $u_{ij} \in (0, 1)$; if the distances from a data point to the centroids $\mathbf{m}_i$ are similar, then $0 < u_{ij} < 1$, and u_{ij} gets close to 0 or 1 as λ decreases to 0.

Theorem 3.2. *With $\lambda > 0$, we fix the values of $\mathbf{m}_i$, $1 \leq i \leq k$ or assume that they have the values at a local minimum. Then, we have the followings for $1 \leq j \leq n$, depending on the value of k:*

- *$k=2$: $\left|\|\mathbf{x}_j-\mathbf{m}_1\|^2-\|\mathbf{x}_j-\mathbf{m}_2\|^2\right| < 2\lambda$, if and only if $0 < u_{1j}, u_{2j} < 1$.*
- *$k \geq 3$: $\sum_{i=1}^{k-1}\left(\|\mathbf{x}_j-\mathbf{m}_{(k)}\|^2-\|\mathbf{x}_j-\mathbf{m}_{(i)}\|^2\right) < 2\lambda$ if and only if $0 < u_{ij} < 1$ for $1 \leq i \leq k$. In this case, $\|\mathbf{x}_j-\mathbf{m}_i\|^2$ are listed in ascending order. That is, $\|\mathbf{x}_j-\mathbf{m}_{(1)}\|^2 \leq \ldots \leq \|\mathbf{x}_j-\mathbf{m}_{(k)}\|^2$.*

Proof. The objective can be written as $\sum_{j=1}^n\left(\sum_{i=1}^k u_{ij}\|\mathbf{x}_j-\mathbf{m}_i\|^2+\lambda\sum_{i=1}^k u_{ij}^2\right)$ and it is separable with respect to j. Hence it suffices to consider the minimum of

$$\sum_{i=1}^k u_i\|\mathbf{x}-\mathbf{m}_i\|^2+\lambda\sum_{i=1}^k u_i^2,$$

by dropping the index j. We further denote $c_i=\|\mathbf{x}-\mathbf{m}_i\|^2$ for simplicity. Now we have the following quadratic function:

$$\lambda\sum_{i=1}^k u_i^2+\sum_{i=1}^k c_iu_i. \tag{5}$$

$k=2$: By substituting $u_2=1-u_1$, the objective becomes $2\lambda u_1^2+(c_1-c_2-2\lambda)u_1+c_2+\lambda$, subject to $0 \leq u_1 \leq 1$. The condition for the vertex $u_1^*=\frac{2\lambda-c_1+c_2}{4\lambda}$ to be in the open interval (0, 1) is equivalent to $-2\lambda < c_1-c_2 < 2\lambda$ or $|c_1-c_2| < 2\lambda$.

$k \geq 3$: Minimizaing the function (5) is equivalent to the following convex quadratic optimization problem:

$$\begin{array}{l}\min_{\mathbf{u}} \frac{1}{2}\left\|\mathbf{u}+\frac{\mathbf{c}}{2\lambda}\right\|^2 \\ \text{subject to} \quad u_i \geq 0, \quad \sum_{i=1}^k u_i=1.\end{array}$$

Let $c_{(i)}$, $i=1,\ldots,k$ be the ascending ordered sequence of c_i, i.e., $c_{(1)} \leq \cdots \leq c_{(k)}$. By applying a projection technique onto simplex [14,15], the optimal $\mathbf{u}^*$ has the closed form

$$u_i^*=\max\left\{-\frac{c_i}{2\lambda}+\beta, 0\right\}, \quad i=1,\ldots,k$$

where

$$\beta=\frac{1}{p}\left(1+\sum_{i=1}^p\frac{c_i}{2\lambda}\right)$$

and p is the largest index such that $-\frac{c_\ell}{2\lambda}+\frac{1}{\ell}\left(1+\sum_{i=1}^\ell\frac{c_i}{2\lambda}\right) > 0$, $\ell=1,\ldots,k$. By observing that $u_1^* \geq \cdots \geq u_k^*$, we must have $p=k$ in order for u_k^* to be positive. Hence we have

$$-\frac{c_k}{2\lambda}+\frac{1}{k}\left(1+\sum_{i=1}^k\frac{c_i}{2\lambda}\right) > 0,$$

equivalently,

$$\sum_{i=1}^{k-1}(c_{(k)}-c_{(i)}) < 2\lambda.$$

□

In (4), a data point is allowed to have non-zero probabilities to multiple clusters. Also, the data points are enforced to belong to one or more clusters (i.e., outliers are not allowed) due to the last equality constraint. To allow outliers, one may relax the equality constraint $\sum_{i=1}^k u_{ij}=1$ by an inequality constraint $\sum_{i=1}^k u_{ij} \leq 1$ to give a chance for the sum to be zero. However, if we replace the equality constraint with the inequality constraint, a global optimizer U is the matrix with all zeros since the corresponding objective value is zero. Thus, we add a penalty term which avoids introducing too many outliers as follows: With $\lambda \geq 0$ and $\nu > 0$,

$$\begin{array}{ll}\min_{U,\mathbf{m}} & \sum_{i=1}^k\sum_{j=1}^n u_{ij}\|\mathbf{x}_j-\mathbf{m}_i\|^2+\lambda\sum_{i=1}^k\sum_{j=1}^n u_{ij}^2+\nu\sum_{j=1}^n\left(\sum_{i=1}^k u_{ij}-1\right)^2 \\ & \text{subject to} \quad u_{ij} \geq 0, \quad 1 \leq i \leq k, \ 1 \leq j \leq n, \\ & \qquad \sum_{i=1}^k u_{ij} \leq 1, \quad 1 \leq j \leq n.\end{array} \tag{6}$$

We also call the model (6) SP K-MEANS. If it is necessary to distinguish it with (4), we call SP K-MEANS with outlier detection.

The next theorem shows how outliers are controlled by the parameter ν. The KKT condition of (6) provides a necessary and sufficient condition for outlier; the distance from the center of cluster must be larger than 2ν, which is given as the next theorem.

Theorem 3.3. *We fix the values of* $\mathbf{m}_i$, $1 \le i \le k$ *or assume that they have the values at a local minimum. If* $u_{ij} = 0$ *for* $1 \le i \le k$ *if and only if* $\|\mathbf{x}_j - \mathbf{m}_i\|^2 \ge 2\nu$ *for all* $1 \le i \le k$.

Proof. Similar to the proofs of previous theorems, we consider

$$\sum_{i=1}^{k} c_i u_i + \lambda \sum_{i=1}^{k} u_i^2 + \nu \left(\sum_{i=1}^{k} u_i - 1 \right)^2, \tag{7}$$

by dropping the index j and denoting $c_i = \|\mathbf{x} - \mathbf{m}_i\|^2$. With the corresponding constraints, the KKT conditions of (7) are

$$\begin{aligned} &u_i \ge 0,\ 1 \le i \le k, \quad \textstyle\sum_{i=1}^{k} u_i \le 1, \\ &\mathbf{c} + 2\lambda \mathbf{u} + 2\nu \left(\textstyle\sum_{i=1}^{k} u_i - 1 \right) \mathbf{1} + \alpha \mathbf{1} - \boldsymbol{\beta} = \mathbf{0}, \quad \alpha \ge 0, \quad \boldsymbol{\beta} \in \mathbb{R}_+^k, \\ &\alpha \left(\textstyle\sum_{i=1}^{k} u_i - 1 \right) = 0, \quad \beta_i u_i = 0,\ 1 \le i \le k, \end{aligned}$$

where α and $\boldsymbol{\beta}$ are Lagrange multipliers. By Karush-Kuhn-Tucker theorem [16], $\mathbf{u}^* = \mathbf{0}$ is an optimal solution if and only if there exist α and $\boldsymbol{\beta}$ which, together with $\mathbf{u}^*$, satisfy the KKT condition.

Hence we get

$$\mathbf{c} - \nu \mathbf{1} - 2\boldsymbol{\beta} = \mathbf{0}, \quad \boldsymbol{\beta} \in \mathbb{R}_+^k.$$

Since $\beta_i \ge 0$ for $1 \le i \le k$, we conclude

$$c_i \ge 2\nu, \quad 1 \le i \le k.$$

□

4. Alternating minimization

In this section, we consider a numerical procedure to solve the proposed models (4) and (6). Although those models are nonconvex, they have two block variables and are convex with respect to each block variable. Using this property, we propose an alternating minimization method for solving those models.

To alternatively solve the proposed models, it first solves (4) or (6) with respect to U for fixed $\mathbf{m}$ and then solves (4) or (6) with respect to $\mathbf{m}$ for fixed U. We formally describe this alternating minimization method in Algorithm 1 and establish its convergence on stationary points. In the sequel, we denote the objective function in the model (4) or (6) by $f(U, \mathbf{m})$ and C_U stands for the set of matrices $\{U \mid u_{ij} \ge 0,\ 1 \le i \le k,\ 1 \le j \le n,\ \sum_{i=1}^{k} u_{ij} = 1,\ 1 \le j \le n\}$ for (4) and $\{U \mid u_{ij} \ge 0,\ 1 \le i \le k,\ 1 \le j \le n,\ \sum_{i=1}^{k} u_{ij} \le 1,\ 1 \le j \le n\}$ for (6).

Algorithm 1 Alternating Minimization.

```
Update $U^{\ell+1}$ and $\mathbf{m}^{\ell+1}$ from $U^{\ell}$ and $\mathbf{m}^{\ell}$:

1. $U^{\ell+1} = \arg\min_{U \in C_U} f(U, \mathbf{m}^{\ell})$.

2. $\mathbf{m}^{\ell} = \arg\min_{\mathbf{m}} f(U^{\ell+1}, \mathbf{m})$.
```

In the first step, the problem is quadratic programming and one may apply any popular solver for it. In our numerical tests, we use 'fmincon' or 'quadprog' functions in the MATLAB optimization toolbox. In the second step, $\mathbf{m}^{\ell}$ is also given by a weighted average:

$$\mathbf{m}_i^{\ell} = \frac{\sum_{j=1}^{n} u_{ij}^{\ell+1} \mathbf{x}_j}{\sum_{j=1}^{n} u_{ij}^{\ell+1}}.$$

Since the problem is a nonconvex optimization problem, our alternating minimization method described in Algorithm 1 may not converge to a global minimum. What we can guarantee is local convergence; Theorem 4.1 shows that every limit point of the sequence generated by Algorithm 1 is a stationary point of the model (4).

In what follows, ι_{C_U} represents the indicator function of the set C_U, i.e., $\iota_{C_U}(U) = 0$ if $U \in C_U$ and $\iota_{C_U}(U) = \infty$ if $U \notin C_U$. The pair $(U^*, \mathbf{m}^*)$ denotes a stationary point of the function $\tilde{f}(U, \mathbf{m}) := f(U, \mathbf{m}) + \iota_{C_U}$. More precisely, $(U^*, \mathbf{m}^*)$ is a point in dom$\tilde{f}$ which satisfies

$$\tilde{f}'((U^*, \mathbf{m}^*); (D_U, d_{\mathbf{m}})) := \liminf_{\epsilon \downarrow 0} \frac{\tilde{f}(U^* + \epsilon D_U, \mathbf{m}^* + \epsilon d_{\mathbf{m}}) - \tilde{f}(U^*, \mathbf{m}^*)}{\epsilon} \geq 0,$$

for any $(D_U, d_{\mathbf{m}}) \in \mathrm{dom}\tilde{f}$.

Theorem 4.1. *Let* $\{(U^\ell, \mathbf{m}^\ell)\}$ *be a sequence generated by Algorithm 1. Then every cluster point of the sequence is a stationary point of the model (4).*

Proof. We first prove the convergence of the values of $\tilde{f}$. For this purpose, we list a few easy consequences: By the definition of f and C_U, f is continuous and ι_{C_U} is lower semicontinuous (lsc). The functions $\tilde{f}(\cdot, \mathbf{m})$ and $\tilde{f}(U, \cdot)$ are strongly convex for given $\mathbf{m}$ and U, respectively. The Algorithm 1 guarantees $\tilde{f}(U^{\ell+1}, \mathbf{m}^\ell) \leq \tilde{f}(U^\ell, \mathbf{m}^\ell)$, and $\tilde{f}(U^{\ell+1}, \mathbf{m}^{\ell+1}) \leq \tilde{f}(U^{\ell+1}, \mathbf{m}^\ell)$. Finally, $\tilde{f}(U, \mathbf{m}) \geq 0$ for any $U \in C_U$. Hence, $\{\tilde{f}(U^\ell, \mathbf{m}^\ell)\}$ converges to some limit and $\{\tilde{f}(U^{\ell+1}, \mathbf{m}^{\ell+1}) - \tilde{f}(U^\ell, \mathbf{m}^\ell)\} \to 0$ and also $\{\tilde{f}(U^{\ell+1}, \mathbf{m}^\ell) - \tilde{f}(U^\ell, \mathbf{m}^\ell)\} \to 0$.

Next, we establish the following convergence to a cluster point. Let $(U^*, \mathbf{m}^*)$ be any cluster point of $\{(U^\ell, \mathbf{m}^\ell)\}$. Since $\tilde{f}$ is lsc, we have $\tilde{f}(U^*, \mathbf{m}^*) \leq \lim_{\ell \to \infty} \tilde{f}(U^\ell, \mathbf{m}^\ell) < \infty$, so $(U^*, \mathbf{m}^*) \in \mathrm{dom}\tilde{f}$. Now, we claim that for any subsequence

$$\{(U^\ell, \mathbf{m}^\ell)\}_{\ell \in \mathcal{K} \subseteq \{0,1,\ldots\}} \to (U^*, \mathbf{m}^*), \tag{8}$$

the following holds:

$$\{(U^{\ell+1}, \mathbf{m}^\ell)\}_{\ell \in \mathcal{K}} \to (U^*, \mathbf{m}^*). \tag{9}$$

We prove the above statement by contradiction. If it does not hold, then there is an infinite subsequence $\mathcal{K}'$ of $\mathcal{K}$ and a number $\delta > 0$ such that

$$\|U^{\ell+1} - U^\ell\|_2 \geq \delta \quad \text{for all } \ell \in \mathcal{K}'.$$

By further passing to a subsequence if necessary, and using the convergence of $\tilde{f}(U^\ell, \mathbf{m}^\ell) = f(U^\ell, \mathbf{m}^\ell) + \iota_{C_U}(U^\ell)$, there exists a nonzero matrix D which holds for

$$\{(U^{\ell+1} - U^\ell)/\|U^{\ell+1} - U^\ell\|_F\}_{\ell \in \mathcal{K}'} \to D, \tag{10}$$

and a scalar θ with

$$\{f(U^\ell, \mathbf{m}^\ell) + \iota_{C_U}(U^\ell)\}_{\ell \in \mathcal{K}'} \to \theta. \tag{11}$$

Fix any $\epsilon \in [0, \delta]$. We let

$$\hat{U}^* = U^* + \epsilon D, \tag{12}$$

and denote

$$\hat{U}^\ell = U^\ell + \epsilon \frac{(U^{\ell+1} - U^\ell)}{\|U^{\ell+1} - U^\ell\|_F}, \tag{13}$$

for each $\ell \in \mathcal{K}'$. By using (8), (10), and (12), we have the limit

$$(\hat{U}^\ell, \mathbf{m}^\ell) \to (\hat{U}^*, \mathbf{m}^*). \tag{14}$$

We note that for each $\ell \in \mathcal{K}'$, $(U^{\ell+1}, \mathbf{m}^\ell)$ is obtained from $(U^\ell, \mathbf{m}^\ell)$ in the first step of the Algorithm 1. Since $\frac{\epsilon}{\|U^{\ell+1} - U^\ell\|_F} \leq \frac{\epsilon}{\delta} \leq 1$, the point $(\hat{U}^\ell, \mathbf{m}^\ell)$ is on the line segment joining $(U^\ell, \mathbf{m}^\ell)$ and $(U^{\ell+1}, \mathbf{m}^\ell)$. This together with $\tilde{f}(U^{\ell+1}, \mathbf{m}^\ell) \leq \tilde{f}(U^\ell, \mathbf{m}^\ell)$ and the convexity of $\tilde{f}(U, \mathbf{m}^\ell)$ with respect to U implies that

$$\tilde{f}(\hat{U}^\ell, \mathbf{m}^\ell) \leq \tilde{f}(U^\ell, \mathbf{m}^\ell) \quad \text{for all } \ell \in \mathcal{K}'. \tag{15}$$

Since $\tilde{f}$ is lsc, (14) and (15) imply $(\hat{U}^\ell, \mathbf{m}^\ell) \in \mathrm{dom}\tilde{f}$.

Also, (15) combined with (11) yields

$$\lim_{\ell \to \infty, \ell \in \mathcal{K}'} \sup\{f(\hat{U}^\ell, \mathbf{m}^\ell) + \iota_{C_U}(\hat{U}^\ell)\} \leq \theta. \tag{16}$$

The fact that $\{\tilde{f}(U^{\ell+1}, \mathbf{m}^\ell) - \tilde{f}(U^\ell, \mathbf{m}^\ell)\}_{\ell \in \mathcal{K}'} \to 0$ and the definition of $\tilde{f}$ imply that

$$\{f(U^{\ell+1}, \mathbf{m}^\ell) + \iota_{C_U}(U^{\ell+1}) - f(U^\ell, \mathbf{m}^\ell) - \iota_{C_U}(U^\ell)\}_{\ell \in \mathcal{K}'} \to 0,$$

and so (11) implies

$$\{f(U^{\ell+1}, \mathbf{m}^\ell) + \iota_{C_U}(U^{\ell+1})\}_{k \in \mathcal{K}'} \to \theta. \tag{17}$$

Let $\gamma = f(\hat{U}^*, \mathbf{m}^*) + \iota_{C_U}(\hat{U}^*) - \theta$. Since f and ι_{C_U} are lsc, we have from (14) and (16) that $\gamma \leq 0$. We proceed to show that $\gamma = 0$. If it is not true, then $\gamma < 0$. In this case, the continuity of f, (14), (16), and (17) imply that

$$f(\hat{U}^*, \mathbf{m}^\ell) + \iota_{C_U}(\hat{U}^*) \leq f(U^{\ell+1}, \mathbf{m}^\ell) + \iota_{C_U}(U^{\ell+1}) + \frac{\gamma}{2},$$

for all sufficiently large $\ell \in \mathcal{K}'$. Equivalently,

$$\tilde{f}(\hat{U}^*, \mathbf{m}^\ell) \leq \tilde{f}(U^{\ell+1}, \mathbf{m}^\ell) + \frac{\gamma}{2}.$$

This contradicts to the fact that $U^{\ell+1}$ is the optimal solution of the problem $\min_U \{f(U,\mathbf{m}^\ell) \mid C_U\}$ in the first step of the Algorithm 1. Hence $\gamma = 0$ and so

$$f(\hat{U}^*,\mathbf{m}^*) + \iota_{C_U}(\hat{U}^*) = \theta.$$

Since the choice of ϵ is arbitrary, we have

$$f(U^* + \epsilon D,\mathbf{m}^*) + \iota_{C_U}(U^* + \epsilon D) = \theta, \quad \forall\, \epsilon \in [0,\delta].$$

The above equality implies that $\tilde{f}(U^* + \epsilon D,\mathbf{m}^*)$ is constant (and finite) for all $\epsilon \in [0, \delta]$. This contradicts to the fact that $\tilde{f}(U,\mathbf{m})$ is strongly convex with respect to U. Hence (9) holds.

Now, we continue to show that

$$f(U^*,\mathbf{m}^*) + \iota_{C_U}(U^*) \le f(U,\mathbf{m}^*) + \iota_{C_U}(U) \quad \forall\, U. \tag{18}$$

By the first step of the Algorithm 1, we have

$$f(U^{\ell+1},\mathbf{m}^\ell) + \iota_{C_U}(U^{\ell+1}) \le f(U,\mathbf{m}^\ell) + \iota_{C_U}(U) \quad \forall\, \ell \in \mathcal{K}. \tag{19}$$

Then, for any fixed $U \in C_U$, the first term on the right-hand side of (19) is finite for all $\ell \in \mathcal{K}$. Passing to the limit as $\ell \to \infty$, $\ell \in \mathcal{K}$, and using the continuity of f, the continuity of ι_{C_U} on C_U, (9), and (19), we conclude that (18) holds.

By a similar argument with replacing $\{(U^\ell,\mathbf{m}^\ell)\}_{\ell\in\mathcal{K}}$ by $\{(U^{\ell+1},\mathbf{m}^\ell)\}_{\ell\in\mathcal{K}}$, we also drive that

$$\{(U^{\ell+1},\mathbf{m}^{\ell+1})\}_{\ell\in\mathcal{K}} \to (U^*,\mathbf{m}^*),$$

and

$$f(U^*,\mathbf{m}^*) \le f(U^*,\mathbf{m}) \quad \forall\, \mathbf{m}. \tag{20}$$

For any $(D_U, d_{\mathbf{m}})$, (18) and (20) yield that

$$\tilde{f}'((U^*,\mathbf{m}^*);(D_U,0)) = \langle \nabla_U f(U^*,\mathbf{m}^*),\ (D_U,0)\rangle + \iota'_{C_U}(U^*;D_U) \ge 0$$

and

$$\tilde{f}'((U^*,\mathbf{m}^*);(0,d_{\mathbf{m}})) = \langle \nabla_{\mathbf{m}} f(U^*,\mathbf{m}^*),\ (0,d_{\mathbf{m}})\rangle \ge 0.$$

Finally, by using the above inequalities, we obtain that

$$\begin{aligned}
&\tilde{f}'((U^*,\mathbf{m}^*);(D_U,d_{\mathbf{m}}))\\
&= \langle \nabla f(U^*,\mathbf{m}^*),\ (D_U,d_{\mathbf{m}})\rangle + \liminf_{\epsilon\downarrow 0} \frac{\iota_{C_U}(U^*+\epsilon D_U) - \iota_{C_U}(U^*)}{\epsilon}\\
&= \Big(\langle \nabla_U f(U^*,\mathbf{m}^*),\ (D_U,0)\rangle + \iota'_{C_U}(U^*;D_U)\Big) + \langle \nabla_{\mathbf{m}} f(U^*,\mathbf{m}^*),\ (0,d_{\mathbf{m}})\rangle\\
&\ge 0.
\end{aligned}$$

Therefore $(U^*, \mathbf{m}^*)$ is a stationary point of $\tilde{f}$, and thus, that of the model (4). □

We refer to [17] for the convergence analysis of the alternating minimization method applied to solve more general nonconvex and nondifferentiable minimization problems.

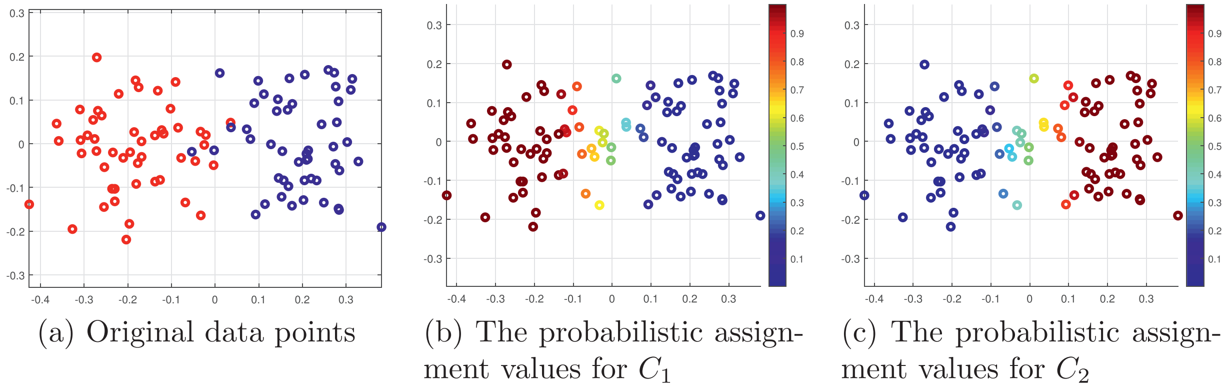
The next theorem establishes that every limit point of the sequence generated by Algorithm 1 is a stationary point of the model (6). Since the proof is similar to that of Theorem 4.1, we omit the proof.

Theorem 4.2. *Let $\{(U^\ell, \mathbf{m}^\ell)\}$ be the sequence generated by Algorithm 1, then every cluster point of the sequence is a stationary point of the model (6).*

5. Experimental results

We demonstrate the characteristics and soundness of our proposed methods on both synthetic and real-world datasets. We first analyze the output of SP K-MEANS on synthetic datasets. We compare the performance of SP K-MEANS with the standard k-means, fuzzy k-means, and NEO-K-Means algorithms on real-world datasets.

In Fig. 1, we create 100 data points in two-dimensional space using two Gaussian distributions. Fig. 1(a) shows the created data points where each color encodes the two ground-truth clusters. In Fig. 1(b)&(c), we represent the value of each u_{ij} $(i = 1,\dots,100,\ j = 1,2)$ for the two clusters derived by the model (4) with $\lambda = 0.05$. That is, Fig. 1(b) represents the probabilistic assignment values of the data points to the first cluster and Fig. 1(c) represents the values for the second cluster. In these figures, red colored data points indicate that those data points have high probability values of being assigned to the corresponding cluster whereas blue colored data points indicate that those data points have low probability values of being assigned to the corresponding cluster. For the data points whose probability values range from 0.7 to 0.3, the colors are gradually changed from orange, yellow, green, and sky blue. From Fig. 1(b), we see that our model identifies the left cluster by assigning high probability values to about half of the data points. Also, the data points which are placed on the middle are colored by orange, yellow, green, and sky blue from left to right indicating that these data points are considered to be the boundary between the two clusters, and thus they have non-zero probabilities to both of the clusters. Among these



(a) Original data points (b) The probabilistic assignment values for C_1 (c) The probabilistic assignment values for C_2

Fig. 1. Data points from two Gaussian distributions and the clustering results derived by the model (4) with $\lambda = 0.05$. Our model identifies the two ground-truth clusters while assigning non-zero probabilities to the boundary data points.

boundary data points, the data points which are placed closer to the left have higher probability values of being assigned to the left cluster annotated by orange and yellow colors. Similarly, in Fig. 1(c), our model successfully identifies the right cluster and assigns non-zero probability values to the boundary data points.

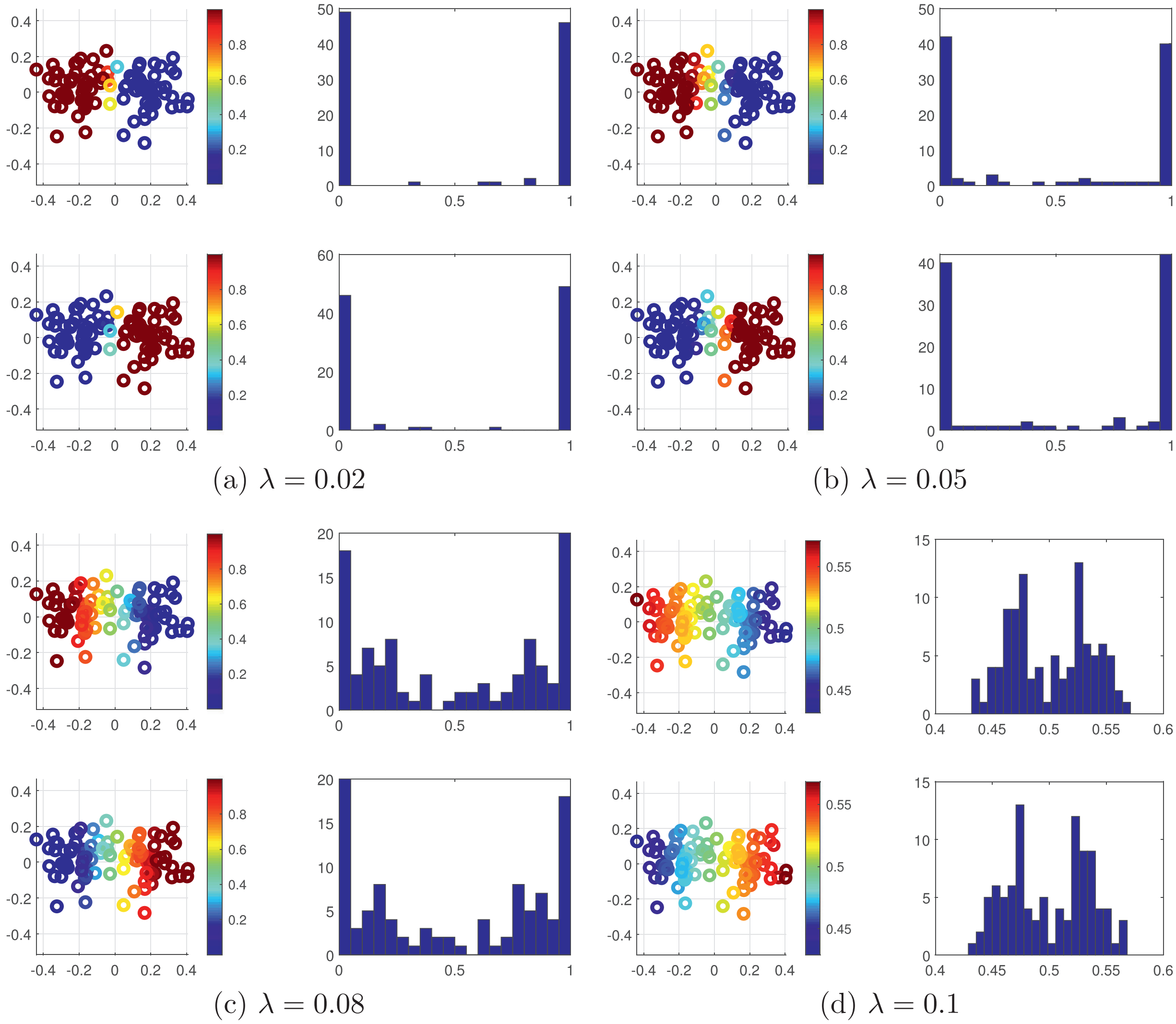
In Fig. 2, we vary the λ values in our model (4). We consider $\lambda = 0.02$, $\lambda = 0.05$, $\lambda = 0.08$, and $\lambda = 0.1$. The histograms show the distributions of the probability values for each cluster. As described in Section 3, as λ becomes close to zero, the entries u_{ij} in the assignment matrix become closer to zeros or ones, i.e., it is preferable to have discrete values in U if λ becomes close to zero (recall Theorem 3.2). From Fig. 2, we see that as λ decreases, the probabilistic assignment values become closer to zeros or ones. On the other hand, as λ increases, the probabilistic assignment values become closer to around 0.5. We note that our model (4) reasonably discovers the two ground-truth clusters with all tested λ values while a different λ value yields a different distribution of the probabilistic assignment values.

To test SP k-means with outlier detection (6), we create 100 data points from two Gaussian distributions and add two outliers as shown in Fig. 3(a) where the two black data points are outliers. Fig. 3(b)&(c) represent the probabilistic assignment values for the first and the second cluster, respectively. We set $\lambda = 0.05$ and $\nu = 0.5$ in our model (6). We note that our model identifies the two ground-truth clusters while assigning non-zero probabilities to both of the clusters for the data points which are placed on the boundary of the clusters. More importantly, the outliers are colored by dark blue in both Fig. 3(b)&(c) meaning that these data points have almost zero probabilities to both of the clusters. This indicates that our model (6) successfully identifies the outliers.

In Fig. 4, we test different ν values in our model (6). Given a fixed $\lambda = 0.05$, we use $\nu = 0.1$, $\nu = 0.5$, and $\nu = 1$. As described in Section 3, in our model (6), we consider a data point to be an outlier if the sum of distances from the data point to the k clusters is proportionally larger than ν. Therefore, the ν value can be considered to be a threshold to determine an outlier; a small ν tends to produce more outliers whereas a large ν tends to produce less outliers (recall Theorem 3.3). In Fig. 4(a)&(b), we see that the two ground-truth outliers have zero probabilities to both of the clusters. When we compare Fig. 4(a)&(b), the data points which are relatively far from the cluster centers have lighter color meaning that they have lower probability values in Fig. 4(a) than Fig. 4(b). Also, when we look at the histograms, the highest probability values are concentrated on around 0.6 in Fig. 4(a) whereas the highest probability values are concentrated on around 0.9 in Fig. 4(b). Intuitively, this is because as ν increases, the model tends to consider more data points to be inliers, and thus is likely to be more confidently assign the data points to their closest clusters, resulting in producing high probability values. In Fig. 4(c), we see that the left corner outlier has a low probability of being assigned to the left cluster and the right corner outlier has a low probability of being assigned to the right cluster when $\nu = 1$. With a large ν, the model (6) becomes generous to outliers and allowing the outliers to have non-zero probability values to their closest clusters.

Now, we test the performance of SP k-means on real-world datasets. We use three real-world datasets (`vision1`, `vision2`, and `music`) where the ground-truth clusters are defined. Since we have the ground-truth clusters, we can compute the F1 scores and the pairwise F1 scores which are also used in [11] to measure the similarity between the ground-truth clusters and the algorithmic clusters. A higher F1 (and pairwise F1) score indicates a better clustering performance. We consider the standard k-means algorithm [12], fuzzy k-means algorithm [2], and the NEO-K-Means algorithm [11] as baseline methods.

We first consider `vision1` dataset where we have a set of 9751-dimensional feature vectors for 390 images. On this dataset, two ground-truth overlapping clusters are defined. Similarly, on `vision2` dataset, there are 915 images, each of which is represented by a 9751-dimensional vector, and seven ground-truth clusters are defined. More details are provided in [11]. We also consider `music` dataset from [18] where there are 593 feature vectors, each of which represents a song by a 72-dimensional feature vector. In this dataset, each song is labelled by the emotions expressed in the song. There are six different pre-defined emotions: amazed-surprised, happy-pleased, relaxing-calm, quiet-still, sad-lonely, and angry-



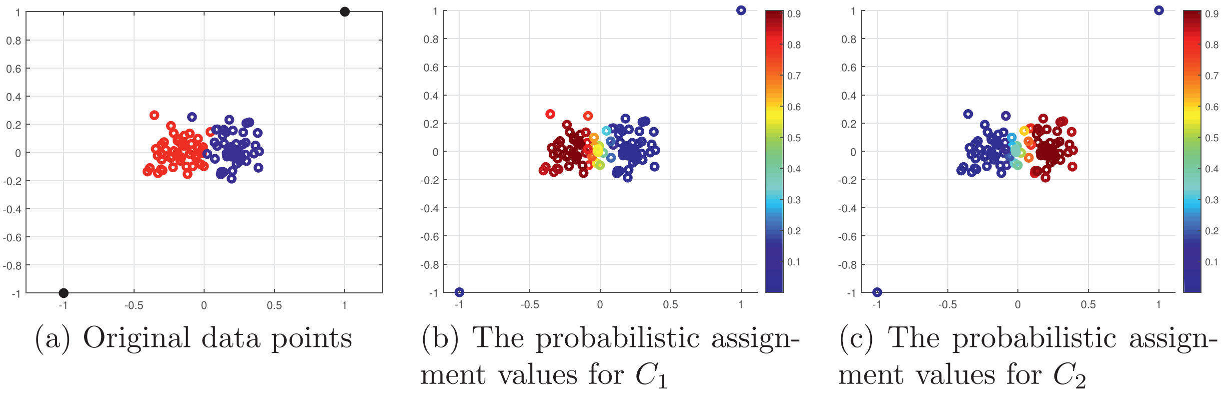
(a) $\lambda = 0.02$

(b) $\lambda = 0.05$

(c) $\lambda = 0.08$

(d) $\lambda = 0.1$

Fig. 2. SP K-MEANS (4) with different λ values. A different λ value yields a different distribution of the probabilistic assignment values.

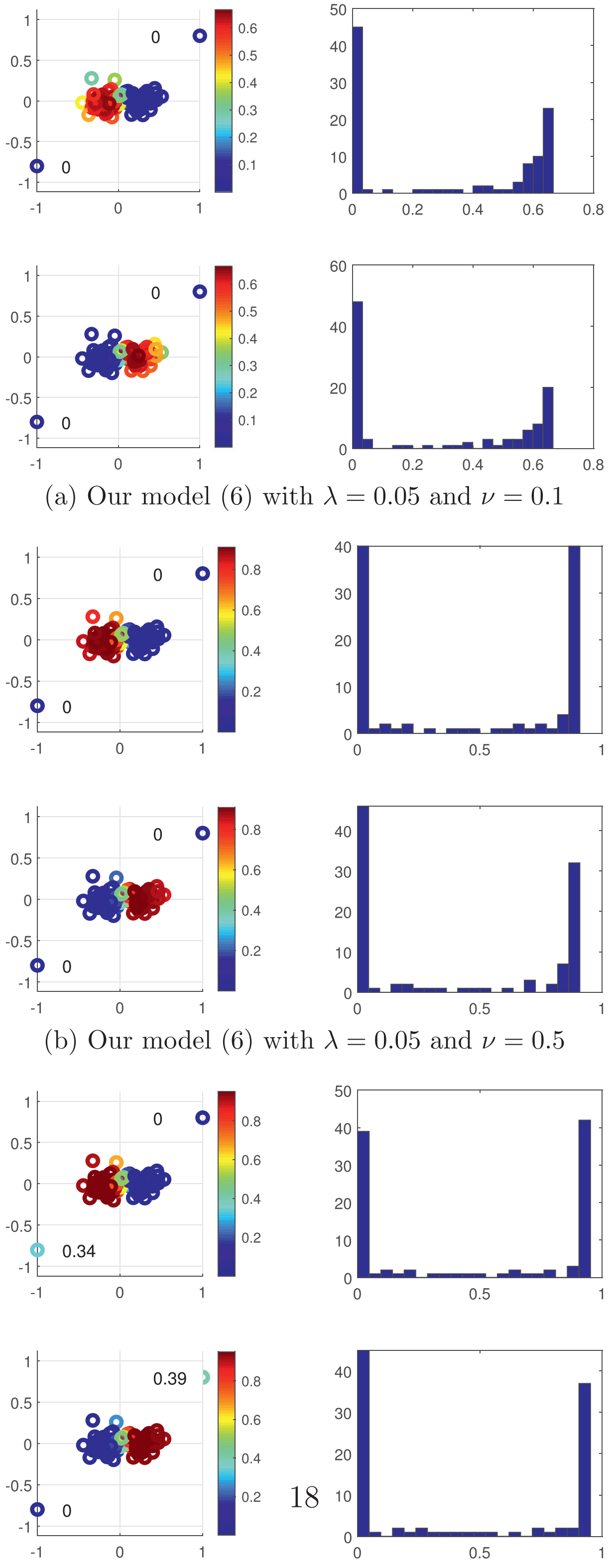


(a) Original data points

(b) The probabilistic assignment values for C_1

(c) The probabilistic assignment values for C_2

Fig. 3. Data points from two Gaussian distributions with two outliers and the clustering results derived by the model (6) with $\lambda = 0.05$ and $\nu = 0.5$. Our model identifies the two ground-truth clusters with probabilistic assignments, and also successfully identifies the outliers.

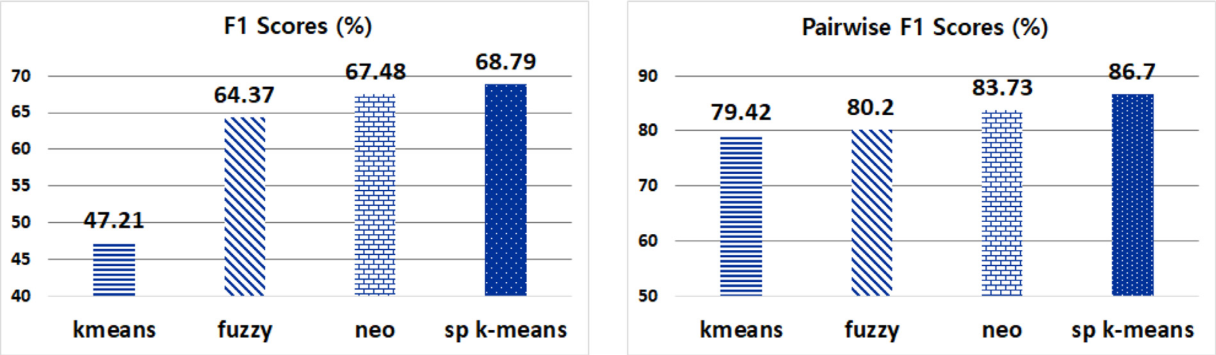


(a) Our model (6) with $\lambda = 0.05$ and $\nu = 0.1$

(b) Our model (6) with $\lambda = 0.05$ and $\nu = 0.5$

(c) Our model (6) with $\lambda = 0.05$ and $\nu = 1$

Fig. 4. SP K-MEANS with outlier detection (6) using different ν values.



(a) F1 Scores

(b) Pairwise F1 Scores

Fig. 5. F1 and pairwise F1 scores on `vision1`. Our model SP k-means achieves the highest F1 and pairwise F1 scores.

Table 1
The best, worst, and average F1 and pairwise F1 scores. On `vision1` dataset, each algorithm returns identical clustering results for five runs.

			kmeans	fuzzy	neo	SP k-means
`music`	F1	best	47.02	46.58	52.41	**54.71**
	(%)	worst	43.20	45.42	48.74	**51.87**
		average	45.04	46.35	51.61	**53.19**
	Pairwise	best	41.65	44.60	60.34	**62.79**
	F1	worst	35.69	42.78	58.46	**61.82**
	(%)	average	38.51	43.14	58.91	**62.35**
`vision1`	F1 (%)	average	47.21	64.37	67.48	**68.79**
	Pairwise F1 (%)	average	79.42	80.20	83.73	**86.70**
`vision2`	F1	best	37.12	N/A	**50.29**	48.99
	(%)	worst	34.37	N/A	**49.66**	46.84
		average	35.47	N/A	**50.02**	47.83
	Pairwise	best	30.78	N/A	52.29	**52.31**
	F1	worst	28.13	N/A	49.39	**51.58**
	(%)	average	29.48	N/A	51.54	**51.76**

aggressive. Each emotion is considered to be a cluster, and a song can involve more than one emotion, which results in overlapping ground-truth clusters.

For the NEO-K-Means algorithm (denoted by neo), we fix $\alpha = \sqrt{k} - 1$. Since SP k-means produces probabilistic assignment values, we should convert them into a hard clustering result. Note that neo makes $(1+\alpha)n$ assignments to yield overlapping clusters where n is the number of data points. To fairly compare the results of SP k-means with neo, we also make $(1+\alpha)n$ assignments for SP k-means by taking the top $(1+\alpha)n$ values of the probabilistic assignment matrix. We fix $\lambda = 0.05$ for SP k-means. For the fuzzy k-means algorithm (denoted by fuzzy), we set the threshold $t = 0.45$ for $k = 2$ and $t = 1/k$ for $k > 2$, and if the probability of a data point to a cluster is greater than t, then we assign the data point to that cluster. Also, we set $r = 1.1$ for the fuzzy K-means since we observe that the fuzzy k-means algorithm assigns all the data points to almost all the clusters if we set $r > 1.1$. We initialize neo, SP k-means, and fuzzy with the result of the k-means algorithm (denoted by kmeans), and run each algorithm five times.

Fig. 5 shows the F1 and pairwise F1 scores (%) on the `vision1` dataset. On this dataset, each algorithm returns identical clustering results for five runs. We see that our algorithm SP k-means achieves the highest F1 and pairwise F1 scores. Also, Table 1 shows the results on the `music`, `vision1`, and `vision2` datasets. For each algorithm, we present the best, the worst, and the average F1 and pairwise F1 scores. On `vision2`, fuzzy assigns the data points to almost all the clusters, resulting in meaningless clusters, so we exclude this result when we compare the clustering results. In Table 1, we see that SP k-means shows the best performance on `music` and comparable results to `NEO` on `vision2`. Therefore, we conclude that our algorithm SP k-means is effective in identifying the ground-truth clusters on real-world datasets. Finally, we present the average runtime of each algorithm in Table 2 where we run each algorithm five times. On the `vision1` dataset, SP k-means is the second fastest algorithm. On the `music` and `vision2` datasets, even though SP k-means requires longer time than the other algorithms to process these datasets, SP k-means produces qualitatively better solutions than the other methods as shown in Table 1.

Table 2
The average runtime of each algorithm in seconds.

	kmeans	fuzzy	neo	SP k-means
`music`	0.0078	1.4129	0.2708	13.5455
`vision1`	0.1508	75.5234	0.7979	0.1756
`vision2`	0.5058	N/A	10.0385	410.2686

6. Conclusion

In this paper, we propose a new clustering model which combines hard and soft clustering. The majority of data which have a clear association follows hard clustering by assigning discrete values and probabilistic values are given to a few data points. To achieve this goal, we properly incorporate the ℓ_1 and ℓ_2 norms. In addition, it can detect outliers, which can cause serious problems in statistical analyses. Through experiments on synthetic data, we show such desired properties and experiments on real-world data demonstrate the usefulness of our model.

Acknowledgments

The authors would like to thank the referee for the helpful comments and constructive suggestions, which helped us to improve the manuscript greatly.

The first author was supported by the National Research Foundation of Korea (NRF) NRF-2016R1A5A1008055, NRF-2016R1D1A1B03931337 and 2019R1F1A1047134. The second author was supported by National Research Foundation of Korea funded by MSIT (2019R1C1C1008956, 2018R1A5A1059921). The third author was supported by the National Research Foundation of Korea (NRF) No. NRF-2016R1D1A1B03934371, NRF-2016R1A5A1008055 and NRF-2019R1F1A1057051.

References

[1] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York-London, 1981.
[3] D.D.S. Chen, M. Saunders, Compressed sensing, IEEE Trans. Info. Theory 52 (2006) 1289–1306.
[4] M. Elad, Sparse and Redundant Representations, Springer, New York, 2010.
[5] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Roy. Statist. Soc. Ser. B 58 (1) (1996) 267–288.
[6] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, 2nd ed., Springer Series in Statistics, Springer, New York, 2009.
[7] L. Ziegelmeier, M. Kirby, C. Peterson, Stratifying high-dimensional data based on proximity to the convex hull boundary, SIAM Rev. 59 (2) (2017) 346–365.
[8] Y.M. Jung, T. Jeong, S. Yun, Non-convex TV denoising corrupted by impulse noise, Inverse Probl. Imaging 11 (4) (2017) 689–702.
[9] J.J. Whang, D.F. Gleich, I.S. Dhillon, Non-exhaustive, overlapping *K*-means, in: Proceedings of the 15th SIAM International Conference on Data Mining, 2015, pp. 936–944.
[10] Y. Hou, J.J. Whang, D.F. Gleich, I.S. Dhillon, Non-exhaustive, overlapping clustering via low-rank semidefinite programming, in: Proceedings of the 21th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2015, pp. 427–436.
[11] J.J. Whang, Y. Hou, D.F. Gleich, I.S. Dhillon, Non-exhaustive, overlapping clustering, IEEE Trans. Pattern Anal. Mach. Intell. 41 (11) (2019) 2644–2659.
[12] S. Lloyd, Least squares quantization in PCM, IEEE Trans. Inf. Theory 28 (2) (1982) 129–137.
[13] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 301–320.
[14] S. Shalev-Shwartz, Y. Singer, Efficient learning of label ranking by soft projections onto polyhedra, J. Mach. Learn. Res. 7 (2006) 1567–1599.
[15] W. Wang, M.A. Carreira-Perpiñán, Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application, 2013, arXiv:1309.1541.
[16] R.T. Rockafellar, Convex Analysis, Princeton Landmarks in Mathematics, Princeton University Press, Princeton, NJ, 1997. Reprint of the 1970 original, Princeton Paperbacks
[17] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (3) (2001) 475–494.
[18] K. Trohidis, G. Tsoumakas, G. Kalliris, I.P. Vlahavas, Multi-label classification of music into emotions, in: Proceedings of the International Conference on Music Information Retrieval, 2008, pp. 325–330.