Statistica Sinica **28** (2018), 1265-1284
doi:https://doi.org/10.5705/ss.202015.0261

SPARSE k-MEANS WITH ℓ_∞/ℓ_0 PENALTY FOR HIGH-DIMENSIONAL DATA CLUSTERING

Xiangyu Chang[1], Yu Wang[2], Rongjian Li[3] and Zongben Xu[1]

[1]*Xi'an Jiaotong University,* [2]*University of California, Berkeley and* [3]*Old Dominion University*

Abstract: One of the existing sparse clustering approaches, ℓ_1-k-means, maximizes the weighted between-cluster sum of squares subject to the ℓ_1 penalty. In this paper, we propose a sparse clustering method based on an ℓ_∞/ℓ_0 penalty, which we call ℓ_0-k-means. We design an efficient iterative algorithm for solving it. To compare the theoretical properties of ℓ_1 and ℓ_0-k-means, we show that they can be explained explicitly from a thresholding perspective based on different thresholding functions. Moreover, ℓ_1 and ℓ_0-k-means are proven to have a screening consistent property under Gaussian mixture models. Experiments on synthetic as well as real data justify the outperforming results of ℓ_0 with respect to ℓ_1-k-means.

Key words and phrases: High-dimensional data clustering, screening property, sparse k-means.

1. Introduction

Clustering is an unsupervised technique for discovering hidden group structures from data sets. It partitions a whole sample set into groups such that each group has its own unique property. The commonly used approaches for clustering include k-means clustering (MacQueen (1967)), hierarchical clustering (Hastie, Tibshirani and Friedman (2009)), model-based clustering (Bishop (2006)) and spectral clustering (Von Luxburg (2007)). In the traditional clustering approaches, all features are treated with equal importance. In fact, only a small portion of features is responsible for intrinsic cluster structures in many applications (Wang et al. (2013)). Those features reflect main characteristics of the data are known as *relevant features*, and the others are usually called *noise features*. The proportion of noise features plays a crucial and negative role for the performance of traditional clustering methods.

Currently, many efforts have been devoted to reduce the influence of noise features on clustering. One common approach is to proceed through dimension reduction, such as principle components analysis (PCA) (Chang (1983)) or non-negative matrix factorization (NMF) (Lee and Seung (1999)), before clustering algorithms are applied. However, existing evidence has shown that these methods do not provide reasonable partitions of the original data (Chang (1983)). Another idea is to perform penalized model-based clustering. It assumes the data matrix is generated from a mixture distribution with unknown parameters. The clusters are uncovered by fitting data into a log-likelihood function with the ℓ_1 penalty (Raftery and Dean (2006); Wang and Zhu (2008); Pan and Shen (2007)). The obvious drawback here is the high computational cost of training the model when the number of features is very large.

Witten and Tibshirani (2010) proposed a framework of sparse clustering that optimizes a weighted cost objective using both the ℓ_1 penalty and ℓ_2 penalty (ℓ_2/ℓ_1 penalty for short). When k-means is selected as the clustering method, they adopted Between-Cluster Sum of Squares (BCSS) as the cost objective and developed a sparse k-means combined with the ℓ_2/ℓ_1 penalty. We call their method ℓ_1-k-means for simplicity, since the ℓ_1 term dominates the final clustering performance compared with the ℓ_2 penalty. Although the performance of ℓ_1-k-means on synthetic data is often good, a considerable portion of noise features is still kept in the final clustering result, as reported in Witten and Tibshirani (2010).

In this paper, we propose a sparse clustering framework for reducing noise features more accurately. Our work starts from the following consensus, proved in (Donoho (2006)), that the ℓ_1 penalty is an optimal convex relaxation of the ℓ_0 penalty. In this paper, therefore, we consider using the ℓ_0 penalty to obtain higher sparsity. Direct application of the ℓ_0 penalty on the sparse clustering framework (Witten and Tibshirani (2010)) results in a solution that cannot be interpreted or explicitly analyzed. To address such challenges, we propose to jointly use the ℓ_∞ and ℓ_0 penalty (ℓ_∞/ℓ_0 penalty for short) for performing clustering. We call this method ℓ_0-k-means when the k-means method is used under our clustering framework. We show the proposed ℓ_0-k-means can be not only explained explicitly from a thresholding perspective, but also analyzed rigorously. In order to justify the effectiveness of our proposed method on clustering, we consioer multiple groups of experiments on synthetic data, and on application data. We show that ℓ_0-k-means exhibits much better noise feature detection capacity than ℓ_1-k-means.

Another important research topic in high-dimensional statistics is analyzing the model behavior when the number of features (variables) grows with the sample size. In the literature (Zhao and Yu (2006); Wainwright (2009)), one finds the variable selection consistency property of the Lasso. Negahban et al. (2012) developed a unified framework for analyzing error bounds of M-estimators with decomposable regularizers, and Fan and Lv (2010) reviewed the techniques about variable selection for penalized regression approaches. Most of these can be categorized as in the supervised learning field. The analysis for the high-dimensional data clustering method, an unsupervised learning method, is still limited (Pan and Shen (2007); Witten and Tibshirani (2010)). We discuss theoretical properties of ℓ_1 and ℓ_0-k-means in this paper. We verify that they can be both interpreted from a thresholding perspective, and that they have screening consistent properties under proper conditions when the data matrix is generated from a high-dimensional Gaussian mixture model.

The rest of the paper is organized as follows. In Section 2, we introduce the existing sparse framework and propose one that includes the ℓ_0-k-means. We give an efficient iterative algorithm to solve for ℓ_0-k-means, and compare the theoretical properties of ℓ_1 and ℓ_0-k-means. In Section 3, we report the finite sample performance of ℓ_0-k-means and other comparable methods on both synthetic data and Allen Developing Mouse Brain Atlas data. We conclude the paper in Section 4. Proofs not included in the main text are presented in the online supplementary material.

2. Sparse Clustering Framework with ℓ_∞/ℓ_0 Penalty

2.1. Existing sparse clustering framework

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the data matrix whose rows $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, $i = 1, \ldots, n$, are samples and columns $\mathbf{X}_j, j = 1, \ldots, p$ are features. Standard k-means clustering groups the data by finding a partition $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$ such that the sum of distances between the empirical mean of each cluster and the corresponding points it contains is minimized. This idea can be generally formulated as an optimization problem,

$$\min_{\mathcal{C},\mu} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mu_k), \tag{2.1}$$

where μ_k is the empirical mean of kth cluster and $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a *dissimilarity measure* satisfying $d(a, a) = 0, d(a, b) \geq 0$ and $d(a, b) = d(b, a)$. The commonly used measure d, is the square of Euclidean distance, $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \sum_{l=1}^{p}(x_{il} - x_{jl})^2$. When *Between-Cluster Sum of Squares (BCSS)* is adopted as the dissimilarity measure function, we can rewrite (2.1) as:

$$\max_{\mathcal{C}} \sum_{j=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \right), \tag{2.2}$$

where $n_k = |C_k|$ is the cardinality of cluster C_k and $d_{ii'j} = (x_{ij} - x_{i'j})^2$. If we take

$$a_j \triangleq \frac{1}{n} \sum_{i,i'}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j}, j = 1, \dots, p, \tag{2.3}$$

then a_j is the jth component of BCSS, which can be considered as a function of only the sample values of the jth feature and the partition $\mathcal{C}$. We use a_j to denote $a_j(\mathcal{C})$ for simplicity. With the formulation (2.3), Witten and Tibshirani (2010) generalized the optimization problem with BCSS (2.2) as

$$\max_{\Theta(\mathcal{C}) \in D} \sum_{j=1}^{p} f_j(\mathbf{X}_j, \Theta(\mathcal{C})), \tag{2.4}$$

where $f_j(\mathbf{X}_j, \Theta(\mathcal{C}))$ is a function that involves only the jth feature of the data, and $\Theta(\mathcal{C})$ is a parameter restricted to a set D. They further defined a *sparse clustering framework*

$$\begin{aligned} &\max_{\mathbf{w}, \Theta(\mathcal{C}) \in D} \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ &\text{s.t.} \quad \|\mathbf{w}\|_2 \leq 1, \ \|\mathbf{w}\|_1 \leq s, \ w_j \geq 0, \ j = 1, \dots, p, \end{aligned} \tag{2.5}$$

where s is a tunning parameter, $\| \cdot \|_2$ is the ℓ_2-norm, $\| \cdot \|_1$ is the ℓ_1-norm, and $\mathbf{w} = (w_1, w_2, \dots, w_p)^\top$ is a weight vector. Here, w_j can be interpreted as the contribution of the jth feature to the objective function (2.5). When they replace $f_j(\mathbf{X}_j, \Theta(\mathcal{C}))$ by a_j as at (2.3), then (2.5) is the ℓ_1-k-means model

$$\begin{aligned} &\max_{\mathcal{C}, \mathbf{w}} \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \right) \\ &\text{s.t.} \quad \|\mathbf{w}\|_2 \leq 1, \ \|\mathbf{w}\|_1 \leq s, \ w_j \geq 0, \ \forall j = 1, \dots, p. \end{aligned} \tag{2.6}$$

Although ℓ_1-k-means have shown excellent performance on a sequence of experiments (Witten and Tibshirani (2010)), they retain some noise features (Wang et al. (2013)). Witten and Tibshirani (2010) gave an example: 60 observations were generated from 3 clusters involving 50 relevant features and 150 noise features, for which ℓ_1-k-means kept all the noise features in the final clustering result. However neither the intuitive explanations on why they can select relevant features nor any theoretical guarantee about their properties have been supplied. In this paper, we propose a new sparse k-means clustering framework to overcome such drawbacks.

2.2. ℓ_0-k-means

The ℓ_1 penalty is commonly replaced by the $\ell_q(0 \le q < 1)$ penalty for sparse modeling problems when more sparsity is needed (Xu et al. (2012); Marjanovic and Solo (2012); Wang et al. (2013)), but this substitution may not be trivial and tractable for sparse clustering. For example, if we use the ℓ_0 penalty instead in (2.5), we have the optimization problem,

$$\begin{aligned} &\max_{\mathbf{w},\Theta(\mathcal{C})} \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ &\text{s.t.} \quad \|\mathbf{w}\|_2 \le 1,\ \|\mathbf{w}\|_0 \le s, w_j \ge 0,\ j = 1, \ldots, p. \end{aligned} \tag{2.7}$$

This model is not easy to analyze or solve since the objective function is no longer convex. Thus we propose to jointly apply the ℓ_∞ and ℓ_0 penalties. We consider the sparse clustering framework

$$\begin{aligned} &\max_{\mathbf{w},\Theta(\mathcal{C})\in D} \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ &\text{s.t.} \quad \|\mathbf{w}\|_\infty \le 1,\ \|\mathbf{w}\|_0 \le s,\ w_j \ge 0,\ j = 1, \ldots, p, \end{aligned} \tag{2.8}$$

where $\|\mathbf{w}\|_\infty = \max_{i=1,2,\ldots,p} |w_j|$ and $\|\mathbf{w}\|_0$ is the number of nonzero components of $\mathbf{w}$.

Similar to ℓ_1-k-means, we define a clustering model by specifying $f_j(\mathbf{X}_j, \Theta(\mathcal{C}))$ to be the a_j at (2.3). Thus, the final objective for the proposed ℓ_0-k-means is

$$\begin{aligned} &\max_{\mathcal{C},\mathbf{w}} \sum_{j=1}^{p} w_j \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \right) \\ &\text{s.t.} \quad \|\mathbf{w}\|_\infty \le 1,\ \|\mathbf{w}\|_0 \le s,\ w_j \ge 0,\ j = 1, \ldots, p. \end{aligned} \tag{2.9}$$

We will show that such ℓ_0-k-means are not only tractable but can be analyzed theoretically. Consider how to solve the ℓ_0-k-means (2.9). The difficulty mainly comes from the existence of two types of variables: the partition variable $\mathcal{C} = \{C_1, \ldots, C_K\}$ that clusters the data samples into K groups, and the weight $\mathbf{w} = (w_1, \ldots, w_p)^\top$ that records the contribution of features. In this paper, we apply the alternative iteration technique to solve ℓ_0-k-means (2.9): we solve $\mathbf{w}$ and $\mathcal{C}$ alternatively by choosing one as the variable and fixing the other. The iterative series is not guaranteed to converge to the global optimum, but the objective function increases monotonically and achieves its maximal value. Since the sample can only be grouped in a finite number of ways and the optimal weights for each fixed partition are unique based on the subsequent analysis, the feasible set of the optimization is finite. Therefore, the algorithm terminates after finite iterations and reaches a local optimum.

The details of the solving procedure of ℓ_0-k-means are in Algorithm 1.

Algorithm 1 ℓ_0-k-means algorithm

Input:

Cluster number K and data matrix $\mathbf{X}$.

Output:

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

1: $w_1^{new} = w_2^{new} = \cdots = w_p^{new} = 1/\sqrt{p}$.

2: Let $\mathbf{w}^{old} = \mathbf{w}^{new}$. Use k-means to find clusters $C_1, C_2, \ldots, C_K$ based on varied distances $w_j^{old} d_{ii'j}$.

3: Fix $C_1, C_2, \ldots, C_K$. Calculate the following optimization problem to obtain $\mathbf{w}^{new}$:

$$\begin{aligned} &\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{a} \\ &s.t. \ \ \|\mathbf{w}\|_\infty \le 1, \|\mathbf{w}\|_0 \le s, w_j \ge 0. \end{aligned} \tag{2.10}$$

4: Repeat step 2 and 3 until

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

Theorem 1. *When the sequence $\{a_j\}_{j=1}^p$ at* (2.3) *satisfies $a_i \ge a_j$ for any $i < j$, an optimal solution of* (2.10) *is given by, for $j = 1, \ldots, p$,*

$$w_j^* = \begin{cases} 1 & j \le \lfloor s \rfloor, \\ 0 & j > \lfloor s \rfloor, \end{cases} \tag{2.11}$$

where $\lfloor s \rfloor$ is the integer part of s.

The solution of (2.10) thus has a closed-form. With the $\{a_j\}_{j=1}^p$ ordered, we assign $w_j = 1$ to the components corresponding to the first $\lfloor s \rfloor$ elements of $\{a_j\}_{j=1}^p$, and $w_j = 0$ to the other elements.

We observe that the standard k-means costs $O(nKp)$ in time complexity and Step 3 of Algorithm 1 costs $O(p\lfloor s \rfloor)$. Thus, the proposed ℓ_0-k-means algorithm is an $O(nKp)$ (if $\lfloor s \rfloor \le nK$) complexity method, the same as the standard k-means. In fact, the condition $\lfloor s \rfloor \le nK$ is easy to satisfy because the number of relevant features is often assumed to be only a small portion of all features in high-dimensional data clustering problems. The ℓ_0-k-means is very efficient in implementation.

2.3. Theory

We analyze the theoretical properties of ℓ_1 and ℓ_0-k-means. For this, assume the data matrix is generated from a high-dimensional Gaussian mixture model. The ℓ_1 and ℓ_0-k-means are interpreted from a thresholding perspective, and then we show that the solutions of ℓ_1 and ℓ_0-k-means have a screening consistent property under mild conditions. We also compare the two models.

Data Generation Model: Suppose each row $\mathbf{x}_i$ of the data matrix $\mathbf{X}$ is i.i.d. from the Gaussian mixture model where

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \phi_{ik} z_{ik}, \tag{2.12}$$

where z_{ik} is a normal random vector with covariance matrix Σ and mean

$$(\vec{v}_k)_j = \begin{cases} \mu_{kj} & j = 1, \ldots, p^*, \\ 0 & j = p^* + 1, \ldots, p, \end{cases} \tag{2.13}$$

and $\phi_{ik} \in \{0, 1\}$ is a binary with $\mathbb{P}(\phi_{ik} = 1) = \pi_k$ and $\sum_{k=1}^{K} \phi_{ik} = 1$ for $k = 1, \ldots, K$. We assume $\sum_k \pi_k \mu_k = 0$ and $\Sigma_{jj} = 1, j = 1, \ldots, p$. We further assume that, for each feature $j = 1, \ldots, p^*$, there exists at least two k and $k' \in \{1, \ldots, K\}$ such that $\mu_{kj} \neq \mu_{k'j}$. With these assumptions, we can ensure that the generated data matrix $\mathbf{X}$ can be distinguished clearly by the first p^* features, the relevant features. Let $\mathcal{C}^* = \{C_1^*, \ldots, C_K^*\}$ be the partition based on $\phi_{ik}, i = 1, \ldots, n, k = 1, \ldots, K$.

Theorem 2. *If the data matrix* $\mathbf{X} = (x_{ij})_{n\times p}$ *is generated according to* (2.12) *and* (2.13)*, then*

$$\mathbb{E}[a_j(\mathcal{C}^*)] = \begin{cases} K - 1 + c_j & 1 \le j \le p^*, \\ K - 1 & \textit{otherwise}, \end{cases} \tag{2.14}$$

where $c_j = n \sum_{k=1}^{K} \pi_k \mu_{kj}^2 - n(\sum_{k=1}^{K} \pi_k \mu_{kj})^2$.

Thus, there is a significant gap between the expectations of relevant and noise features when the data matrix is generated by the Gaussian mixture model. For example, for the jth feature, the gap is $c_j = n \sum_{k=1}^{K} \pi_k \mu_{kj}^2 - n(\sum_{k=1}^{K} \pi_k \mu_{kj})^2 > 0$. Here we used the convexity of function x^2 and the assumption $\mu_{kj} \neq \mu_{k'j}$ for some $k \neq k'$ to obtain the positiveness. The convexity also can be used to prove that the gap c_j grows larger when the K groups are distinguished more clearly on the jth feature.

The ℓ_1-k-means proposed in Witten and Tibshirani (2010) is in fact based on such gaps to distinguish relevant features from noise features. Given an estimated partition $\widehat{\mathcal{C}}$, ℓ_1-k-means define the optimal feature weight

$$\widehat{\mathbf{w}} = \frac{S(\mathbf{a}(\widehat{\mathcal{C}}), \Delta)}{\|S(\mathbf{a}(\widehat{\mathcal{C}}), \Delta)\|_2}, \tag{2.15}$$

where $S(\mathbf{a}, \Delta)_j = \max(a_j - \Delta, 0)$ is the soft thresholding function (Donoho (1995)). From (2.15), we can see that any feature with $a_j < \Delta$ is identified as a noise feature, otherwise it is a relevant feature. Compared with ℓ_1-k-means, Theorem 1 indicates that ℓ_0-k-means use the hard thresholding function (Blumensath and Davies (2008)) to distinguish relevant and noise features. Although ℓ_1 and ℓ_0-k-means both take full advantage of the same gap information to select relevant features, we show their feature selection capacity is different.

Let $\mathcal{C}$ be any partition of the n samples, and its BCSS for feature j be (2.3). By Lemma 1 in the supplementary materials, we know

$$a_j(\mathcal{C}) = -\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\right)^2 + \sum_{k=1}^{K}\left(\frac{1}{\sqrt{|C_k|}}\sum_{i\in C_k} x_{ij}\right)^2. \tag{2.16}$$

We omit the constant term and define the weighted BCSS as

$$\begin{aligned} F(\mathcal{C}, w) &\triangleq \sum_{j=1}^{p} w_j \bar{a}_j(\mathcal{C})^2 \\ &\triangleq \sum_{j=1}^{p} w_j \left\{ a_j(\mathcal{C}) + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\right)^2 \right\} \qquad (2.17) \\ &= \sum_{j=1}^{p} w_j \sum_{k=1}^{K}\left(\frac{1}{\sqrt{|C_k|}}\sum_{i\in C_k} x_{ij}\right)^2. \qquad (2.18) \end{aligned}$$

Our goal is to analyze the screening property of the problem

$$\begin{aligned} \max_{\mathcal{C},\mathbf{w}} \quad & F(\mathcal{C}, \mathbf{w}) \qquad (2.19) \\ s.t. \quad & \mathbf{w} \in \Omega, \end{aligned}$$

where Ω is a constraint set of $\mathbf{w}$.

Definition 1. *The estimated weight* $\widehat{\mathbf{w}}$ *of* (2.19) *has the screening consistent property (SCP) provided*

$$\mathbb{P}(\{1\ldots, p^*\} \subset supp(\widehat{\mathbf{w}})) \to 1, \textit{as } n \to \infty$$

where $supp(\widehat{\mathbf{w}}) = (j|\widehat{w}_j \neq 0, j = 1, \ldots, p)$.

Theorem 3. *Let* $(\widehat{\mathcal{C}}, \widehat{\mathbf{w}})$ *be the optimal solution of* (2.19) *where* $\Omega = \Omega_1 =$

$\{\mathbf{w}|\ \|\mathbf{w}\|_1 \le s, \|\mathbf{w}\|_2 \le 1\}$. *Let* $\sigma_1 = \min_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$, $\sigma_2 = \max_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$. *If* $p^{*2} \le \sigma_1^4/(6400\sigma_2^3 \ln(K))$ *and* $\ln(p) = o(n)$, *with*

$$\frac{\sum_{j=1}^{p^*}\sum_{k=1}^K \pi_k\mu_{kj}^2 - 1/2\sigma_1 p^*}{\sqrt{\sum_{j=1}^{p^*}\left(\sum_{k=1}^K \pi_k\mu_{kj}^2 - 1/2\sigma_1\right)^2}} \le s \le \frac{\sum_{j=1}^{p^*}\sum_{k=1}^K \pi_k\mu_{kj}^2}{\sqrt{\sum_{j=1}^{p^*}\left(\sum_{k=1}^K \pi_k\mu_{kj}^2\right)^2}},$$

we have

$$\mathbb{P}\left(\widehat{\mathbf{w}} \text{ has SCP}\right) \to 1, \text{as } n \to \infty. \tag{2.20}$$

Theorem 4. *With the notation of Theorem* 3, *if* $p^{*2} \le s^2 \le \sigma_1^2/(192\ln(K)\sigma_2)$ *and* $\ln(p) = o(n)$, *then*

$$\mathbb{P}\left(\widehat{\mathbf{w}} \text{ has SCP}\right) \to 1, \text{as } n \to \infty. \tag{2.21}$$

Thus ℓ_1 and ℓ_0-k-means both have the SCP if p^* is small enough and $\ln p = o(n)$. That $\ln p = o(n)$ is considered to be optimal for regularized regression approaches to ultra-high dimensional feature selection problems (see e.g., Zhao and Yu (2006); Wainwright (2009); Fan and Lv (2010)). Although ℓ_1 and ℓ_0-k-means have the same property, their finite sample performance is differs.

3. Experimental Evaluation

In this section, we evaluate and compare the finite sample performance of ℓ_0-k-means with other popular algorithms based on a set of synthetic data and an application to data from the Allen Developing Mouse Brain Atlas.

The ℓ_0-k-means involve a tuning parameter s which controls the number of features selected. Witten and Tibshirani (2010) proposed a strategy to select the tunning parameter s based on the gap statistic (Tibshirani, Walther and Hastie (2001)). We follow their strategy for the proposed ℓ_0-k-means as well. We consider two criteria for comparison. The first is the *Classification Error Rate* (CER) (Witten and Tibshirani (2010); Chipman and Tibshirani (2006)), defined as $CER \triangleq \sum_{i>i'} |1_{\widehat{\mathcal{C}}(i,i')} - 1_{\mathcal{C}^*(i,i')}|/\binom{n}{2}$, where $1_{\mathcal{C}(i,j)}$ is an indicator function to record whether the ith and jth sample are in the same group with respect to partition $\mathcal{C}$. The second criterion is F_1-score, which measures the feature selection accuracy. If

$$\text{precision} = \frac{|(i : w_i \ne 0, \widehat{w}_i \ne 0)|}{|(i : \widehat{w}_i \ne 0)|},$$

and

$$\text{recall} = \frac{|(i : w_i \ne 0, \widehat{w}_i \ne 0)|}{|(i : w_i \ne 0)|},$$

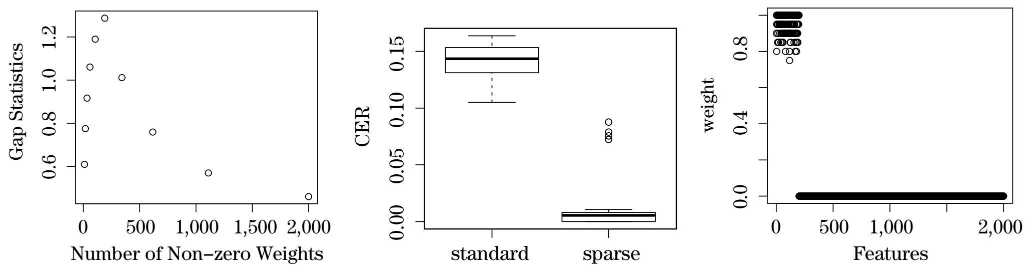

Figure 1. Overview of ℓ_0-k-means.

then F_1-score is the harmonic mean of precision and recall,

$$F_1\text{-score} = 2 \cdot \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

3.1. Evaluation on synthetic data

Four experiments were conducted. The first experiment was to verify that the gap statistic succeeds in selecting an appropriate tuning parameter for ℓ_0-k-means. The second and the third experiments were to compare the performance of ℓ_0-k-means, ℓ_1-k-means, standard k-means, PCA-k-means, and EM algorithm for penalized log likelihood for a Gaussian mixture model with independent or correlated features. In the fourth experiment, we explored the performance of those algorithms for non-Gaussian distributions.

Experiment 1: We constructed 6 clusters, each cluster containing 20 samples with 2,000 features, leading to a data matrix $\mathbf{X}_{120\times 2{,}000}$. Among the 2,000 features, we assumed only the first 200 were relevant features. For the kth cluster, relevant features were sampled from a $\mathcal{N}(0.5 \cdot k, 1)$ and noise features were sampled from $\mathcal{N}(0, 1)$ independently. The data matrix was normalized to have column-wise zero mean before any algorithm was applied. We repeated the sample generation procedure 20 times and report the averaged results based on these 20 trials for ℓ_0-k-means and standard k-means. The results are shown in Figure 1.

From the left subfigure of Figure 1, we can see that the highest gap statistic is achieved when the number of non-zero weights is around 200. This shows the gap statistic to be useful for the selection of tuning parameter for ℓ_0-k-means. The middle subfigure shows that the obtained partition has a significant smaller CER compared with standard k-means. In the right subfigure, we report the average values of estimated weights over 20 trails for each feature. Here the values for relevant features are approximately 1 while those for noise features are close to 0. Gap statistics for ℓ_0-k-means can help the selection of relevant features and

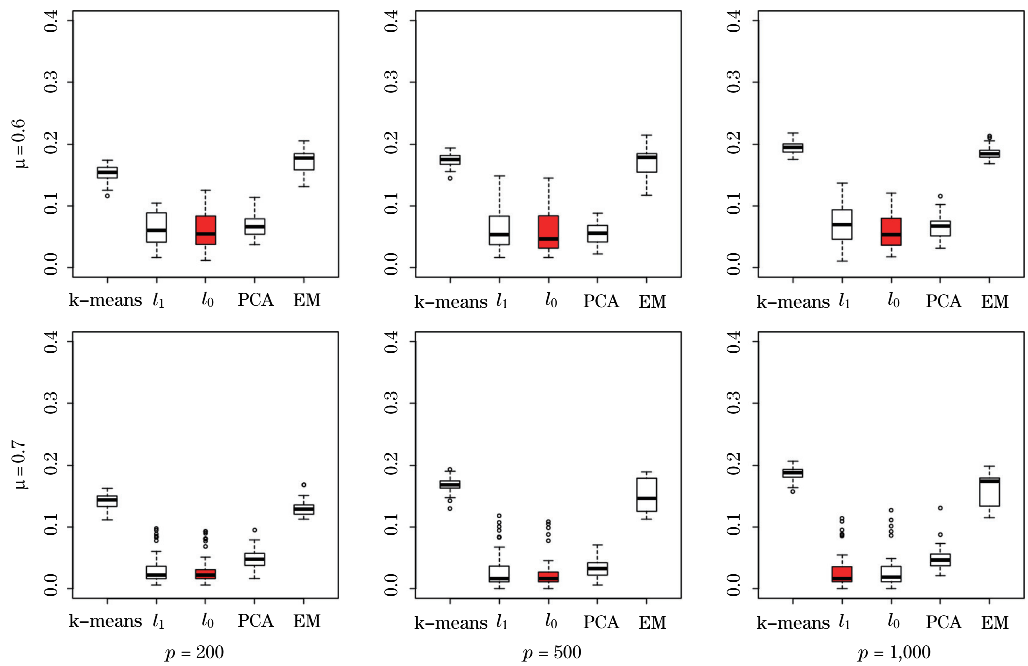

Figure 2. CER Boxplot for Experiment 2.

improve the accuracy of partitions.

Experiment 2: We report the performance of standard k-means, ℓ_0-k-means, ℓ_1-k-means, PCA-k-means (Chang (1983)) (PCA for short), and EM for ℓ_1-penalized log likelihood (Pan and Shen (2007)) (EM for short) when data was generated from a Gaussian mixture model with independent features. We assumed each element x_{ij} in the data matrix was $\mathcal{N}(\mu_{ij}, \sigma_j^2)$ independently, with

$$\mu_{ij} = \begin{cases} a_j\mu & \text{if } i \in C_1,\ j \leq 50, \\ -a_j\mu & \text{if } i \in C_2,\ j \leq 50, \\ 0 & \text{if } i \in C_3,\ \text{ or } j > 50, \end{cases} \tag{3.1}$$

where a_j was chosen randomly from $[0.75, 1.25]$ for each $j = 1, \ldots, 50$, and σ_j was chosen randomly from $[0.75, 1.25]$ for $j = 1, \ldots, p$. Thus, the first 50 features were relevant while the rest were noise. There were 3 clusters and each cluster contained 50 samples, with $\mu = 0.6, 0.7$ and $p =$ 200, 500, 1,000. Each parameter setting was repeated 50 times. The results are reported in Figures 2 and 3.

In Figure 2, ℓ_0-k-means have the best average clustering performance (lowest CER) compared to other algorithms. This can be explained by the superior feature selection performance of ℓ_0-k-means shown in Figure 3. The ℓ_0-k-means,

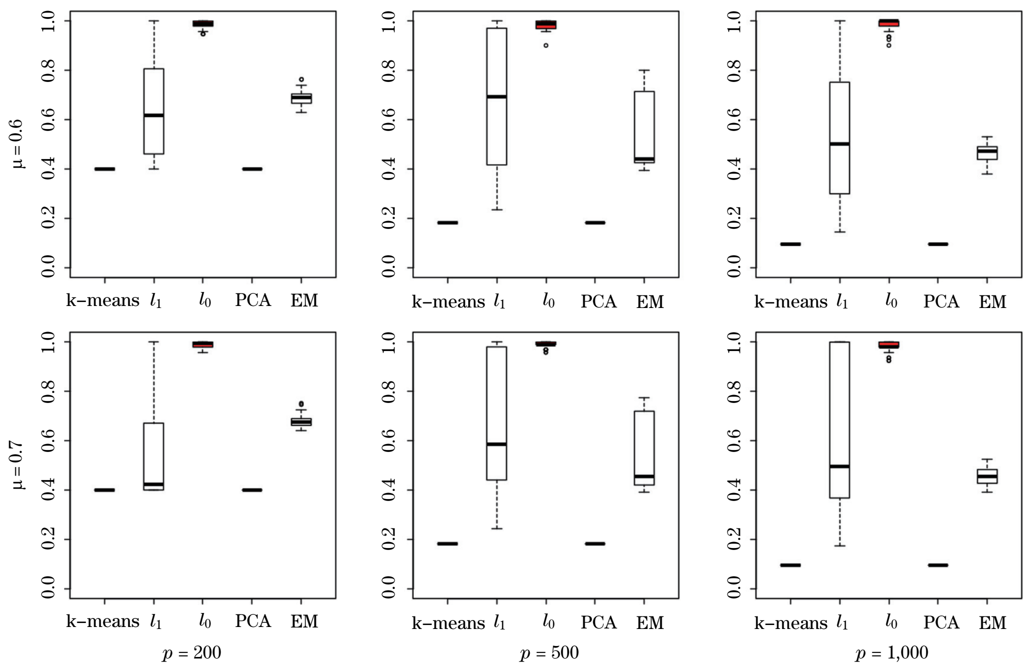

Figure 3. F_1-score Boxplot for Experiment 2.

compared to other algorithms, has F_1-score close to 1 with a small deviation. This may explain why ℓ_0-k-means tend to have lower CERs than the other algorithms.

Experiment 3: Similar to Experiment 2, we report the performance of ℓ_0-k-means when data was generated from a Gaussian mixture model with correlated features. Suppose each sample $\mathbf{x}_i$ was $\mathcal{N}(\mu, \Sigma)$, where the elements Σ_{ij} of Σ were $\Sigma_{ij} = 0.1^{|i-j|}$.

In Figures 4 and 5, it can be seen that the performance of ℓ_0-k-means is quite stable. It always has the highest feature selection F_1-scores and the lowest CER values among the algorithms.

Experiment 4: In this experiment, we extended the Gaussian mixture model to non-Gaussian cases. Experiment settings were identical to those of Experiment 2, except we used the standard log normal distribution $f(x) = k \cdot \mu + a \cdot \exp(\mathcal{N}(0, 1))$ and standard Poisson distribution $f(x) = k \cdot \mu + \text{Poisson}(1)$, with a chosen randomly from $[0.75, 1, 25]$ and $k = 1, \ldots, K$. We took $\mu = 2, 3$ for the log normal distribution and $\mu = 1, 1.5$ for the Poisson distribution. The results are shown in Figures 6 to 9. Here ℓ_0-k-means achieve the best feature selection accuracy.

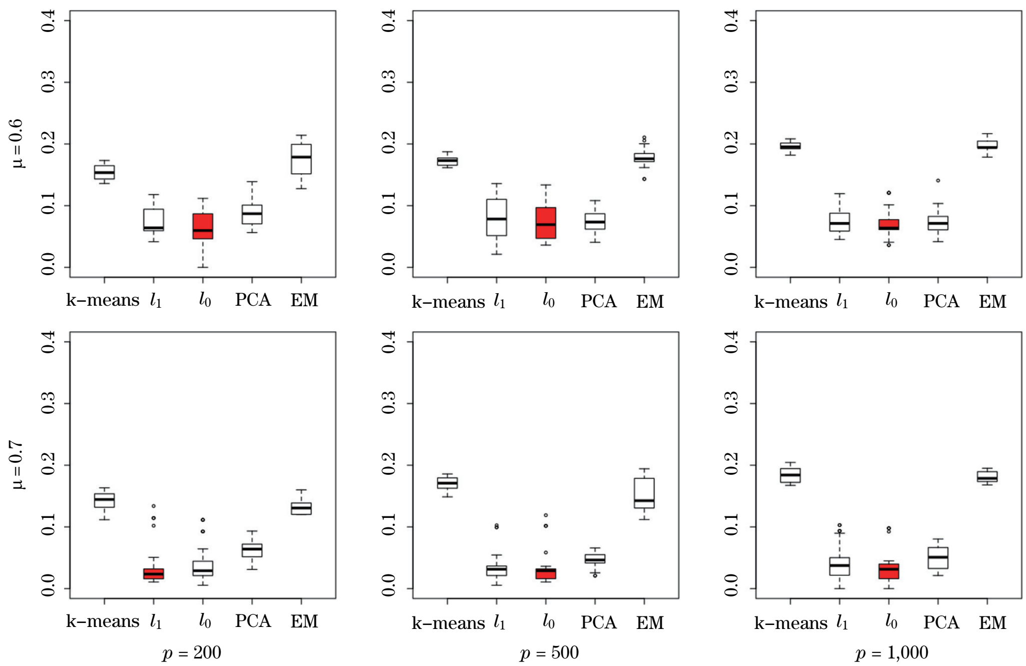

Figure 4. CER Boxplot for Experiment 3.

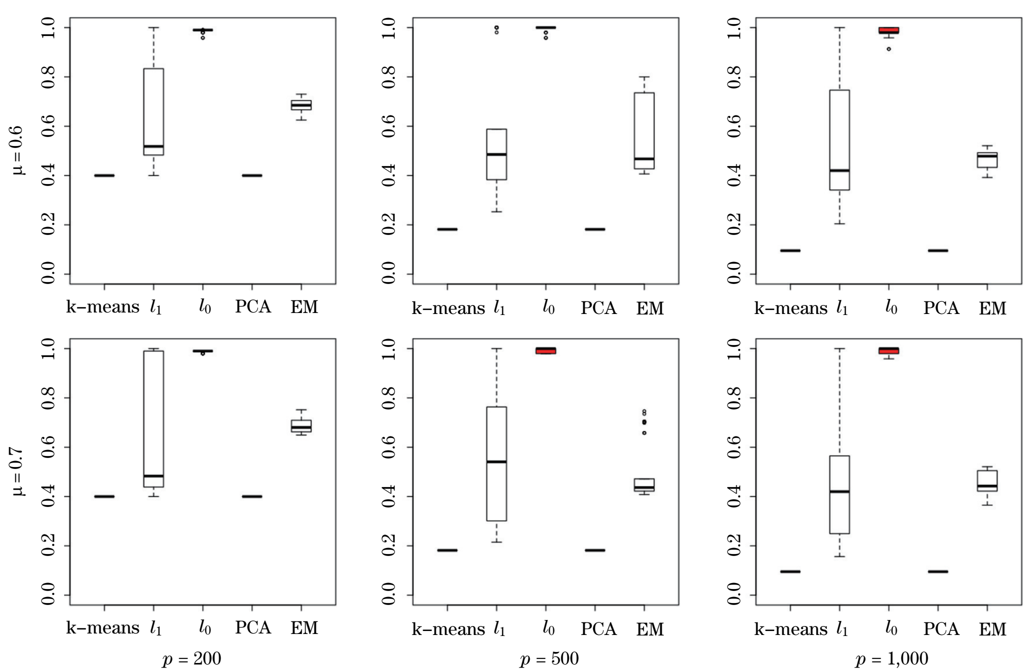

Figure 5. F_1-score Boxplot for Experiment 3.

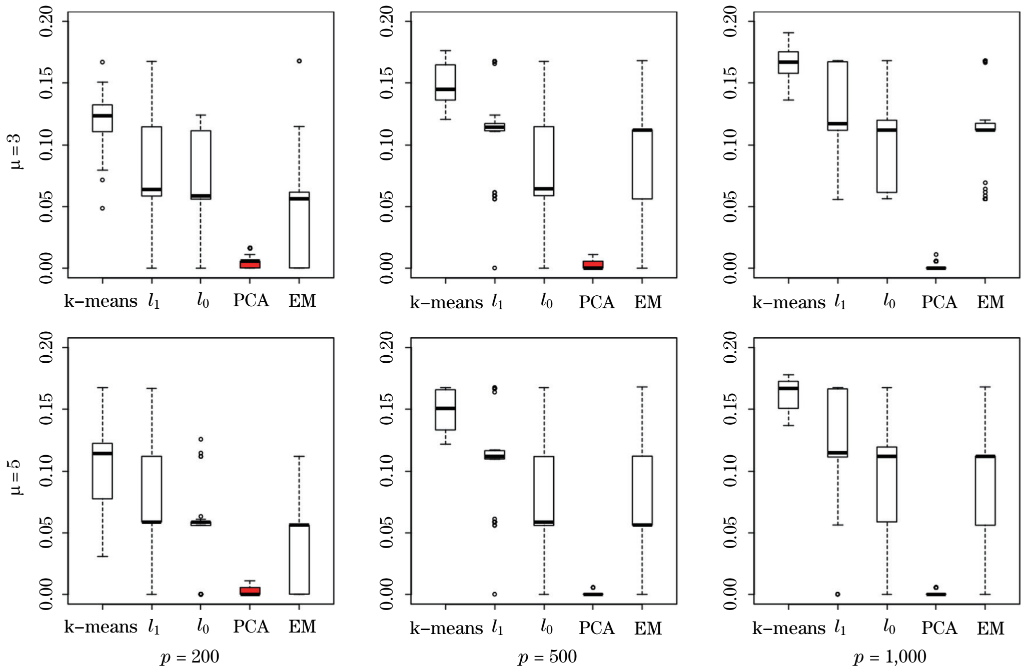

Figure 6. CER Boxplot for Experiment 4 log normal case.

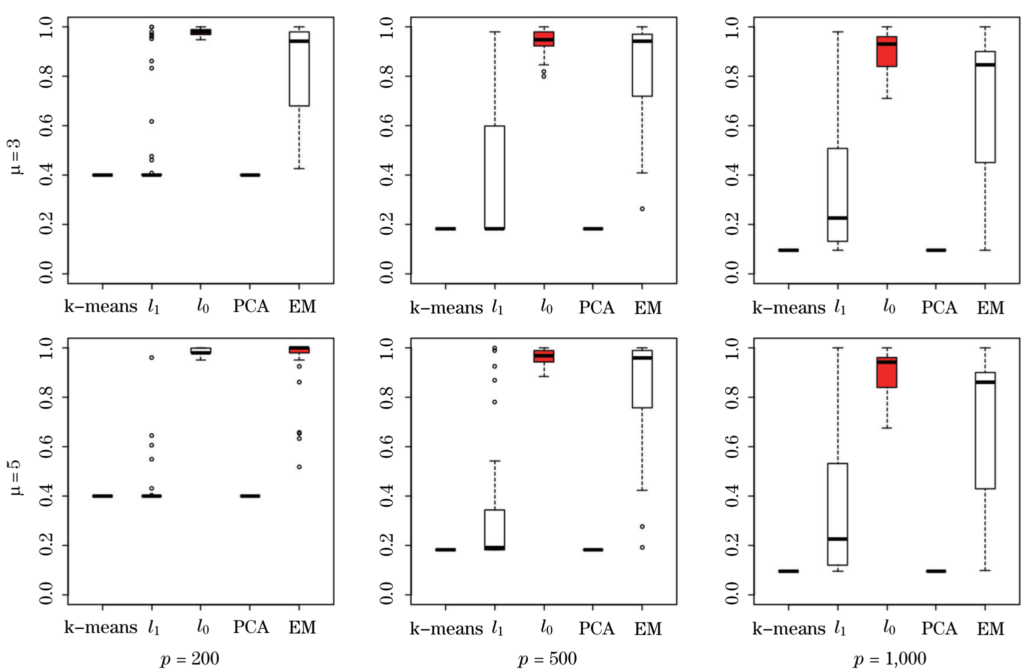

Figure 7. F_1-score Boxplot for Experiment 4 log normal case.

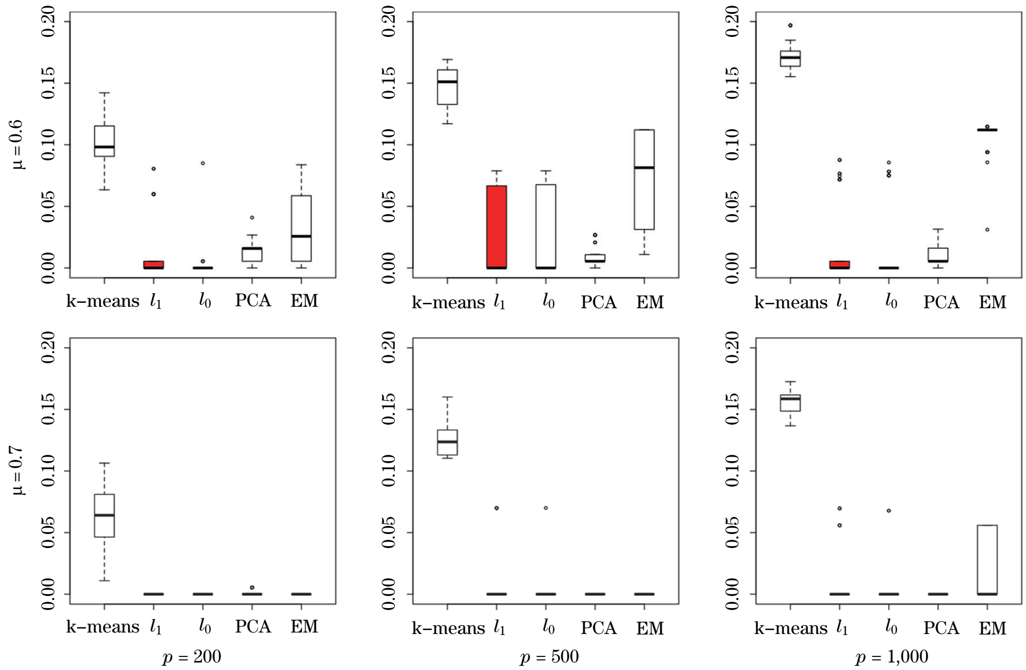

Figure 8. CER Boxplot for Experiment 4 Poisson case.

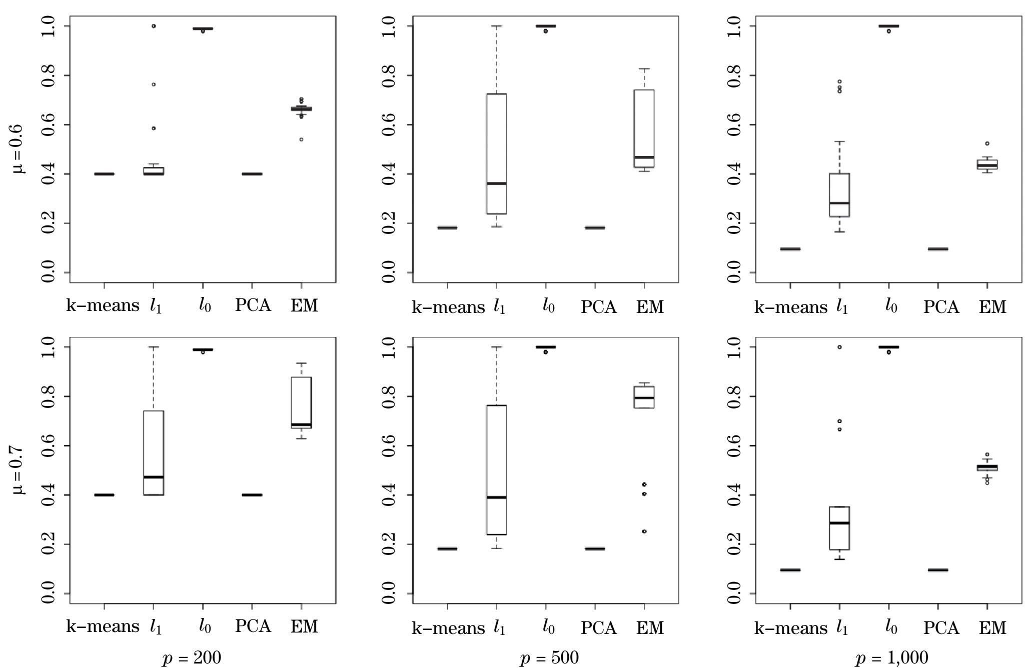

Figure 9. F_1-score Boxplot for Experiment 4 Poisson case.

Table 1. Statistics of mouse brain data at annotation level 3.

Ages	E11.5	E13.5	E15.5	E18.5	P4	P14	P28
Number of genes	1,724	1,724	1,724	1,724	1,724	1,724	1,724
Number of voxels	7,122	13,194	12,148	12,045	21,845	24,180	28,023
Number of regions	20	20	20	20	20	19	20

Figure 10. Selected sample slices of 7 developmental mouse brains with respect to the gene Neurog1.

3.2. Evaluation of the Allen Developing Mouse Brain Atlas

We compared our proposed method with other methods on the Allen Developing Mouse Brain Atlas data (Lein et al. (2007); Li et al. (2015); Wang et al. (2013)). This data set contains *in situ* hybridization gene expression pattern images of a developing mouse brain across 7 developmental ages. The mouse brain is imaged into 3D space with voxels in a regular grid. The expression energy at each voxel for some gene is recorded as a numerical value. Through such operations, 7 data matrices associated with 7 developmental ages are obtained. In these data matrices, rows correspond to brain voxels and columns correspond to genes. With the development of a mouse brain, the rows of energy matrices increase because, as the size of brain grows larger, more and more voxels are needed to stabilize the resolution. The basic statistics of the data are listed in Table 1, and Figure 10 shows the sample slices of 7 developmental mouse brains with respect to the gene *Neurog1*. In fact, each voxel is annotated with a brain region manually, which can be viewed as the ground truth cluster label.

We applied the ℓ_0-k-means, ℓ_1-k-means, standard k-means, PCA-k-means, and EM for ℓ_1-penalized log likelihood (EM for short) to the 7 data matrices. The results, including CER values and feature selection performance, are shown in Tables 2 and 3. From Table 2, we can see that the ℓ_0-k-means, in most cases, outperforms other competitors. Besides the low CER values while using the smallest number of features (i.e., nonzero weights $\mathbf{w}$), another advantage of ℓ_0-

Table 2. The CER values of clustering when the algorithms are applied to Allen Developing Mouse Brain Atlas data.

Ages	E11.5	E13.5	E15.5	E18.5	P4	P14	P28
k-means	0.1610	0.1877	**0.2055**	0.2369	0.3444	0.3628	0.3599
ℓ_1-k-means	0.1662	0.1985	0.2221	0.2425	0.3308	0.3593	0.3470
ℓ_0-k-means	**0.1605**	**0.1842**	0.2259	**0.2358**	**0.3306**	**0.3580**	0.3505
PCA-k-means	0.1654	0.1977	0.2321	0.2682	0.3617	0.3860	0.3650
EM	0.2471	0.2432	0.3045	0.3100	0.4141	0.3707	**0.3419**

Table 3. The NW values of clustering when the algorithms were applied to Allen Developing Mouse Brain Atlas data.

Ages	E11.5	E13.5	E15.5	E18.5	P4	P14	P28
k-means	1,723	1,724	1,724	1,724	1,720	1,724	1,724
ℓ_1-k-means	717	672	659	642	446	224	1,724
ℓ_0-k-means	100	660	100	1,600	199	322	1,068
PCA-k-means	1,723	1,724	1,724	1,724	1,720	1,724	1,724
EM	1,723	1,724	1,724	1,724	1,720	1,724	1,724

k-means is *interpretability.* Apparently the ℓ_0-k-means can eliminate more noise features than other methods. For instance, consider the postnatal stage P14 as differentiation of gene functions is more discriminative at this postnatal stage. We observe that there are few "noisy" genes which have been eliminated by ℓ_0-k-means and included by ℓ_1-k-means. Thus a noisy gene 'Scn4b' is detected by our ℓ_0-k-means method. This gene is highly related to the protein composition of sodium channel beta subunits (Medeiros-Domingo et al. (2007)), is strongly bonded with electrical signal transmission activities in most of types of cells, and it is reasonable to consider features corresponding to this gene as noise; its function is uniformly supportive in the whole brain and using it to distinguish different regions may not be effective. Detecting a feature as noise by ℓ_1-k-means is consistent with the prior knowledge about genes listed in the database of Allen Institute*.

4. Conclusion and Future Work

In this paper, we focus on designing an efficient clustering algorithm for high dimensional data sets. Inspired by the literature of sparse clustering, we allow algorithms to optimize weights of individual features to combine clustering procedures with feature selection. We proposed a new sparse clustering method

*http://www.genecards.org/.

with ℓ_∞/ℓ_0 penalty, called ℓ_0-k-means. They can be efficiently solved by our Algorithm 1. Both ℓ_0-k-means and ℓ_1-k-means have screening consistency under appropriate conditions for Gaussian mixture model, but empirical experiments suggest that ℓ_0-k-means outperform ℓ_1-k-means in feature selection in terms of F_1-score. Extensive experiments were carried out to compare with some other well-known clustering methods.

In the future, we might carry out our work in the following directions. We intend to investigate the possibility of establishing a feature selection consistency property for ℓ_0 and ℓ_1-k-means within the framework of this paper. We mean to extend the current research by going on to other high-dimensional data clustering models, for instance, penalized model-based clustering (Pan and Shen (2007)).

Supplementary Materials

We provide proofs of the theorems in the online supplementary material.

Acknowledgment

We thank the review team - the Editors and the two anonymous reviewers for their careful work and constructive comments, that greatly helped improve the paper. Xiangyu Chang's research is supported by the National Natural Science Foundation of China (Project No. 11401462, 61502342, 61603162) and the China Postdoctoral Science Foundation (Project No. 2015M582630). The corresponding author is Dr. Rongjian Li.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Blumensath, T. and Davies, M. E. (2008). Iterative thresholding for sparse approximations. *J. Fourier. Anal. Appl.* **14**, 629-654.

Chang, W. C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Appl. Stat.*, 267-275.

Chipman, H. and Tibshirani, R. (2006). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics* **7**, 286-301.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **41**, 613-627.

Donoho, D. L. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35**, 617-652.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20**, 101-148.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd Edition. Springer.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788-791.

Lein, E. S., Hawrylycz, M. J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A. F., Boguski, M. S., Brockway, K. S., Byrnes, E. J. and Chen, L. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168-176.

Li, R., Zhang, W., Zhao, Y., Zhu, Z. and Ji, S. (2015). Sparsity learning formulations for mining time-varying data. *IEEE Trans. Knowl. Data Eng.* **27**, 1411-1423.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, No. 14, pp. 281-297).

Marjanovic, G. and Solo, V. (2012). On l_q optimization and matrix completion. *IEEE Trans. Signal Process.* **60**, 5714-5724.

Medeiros-Domingo, A., Kaku, T., Tester, D. J., Iturralde-Torres, P., Itty, A., Ye, B., Valdivia, C., Ueda, K., Canizales-Quinteros, S., Tusié-Luna, M. T. and Makielski, J. C. (2007). SCN4B-encoded sodium channel β_4 subunit in congenital long-QT syndrome. *Circulation* **116**, 134-142.

Negahban, S., Ravikumar, P. K., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci.* **27**, 538-557.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res.* **8**, 1145-1164.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *J. Amer. Statist. Assoc.* **101**, 168-178.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411-423.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* **17**, 395-416.

Wang, Y., Chang, X., Li, R. and Xu, Z. (2013). Sparse k-means with the $\ell_q(0 \leq q < 1)$ constraint for high-dimensional data clustering. In *IEEE 13th International Conference on Data Mining*, 797-806, Dallas, TX: IEEE.

Wang, S. and Zhu, J. (2008). Variable selection for modelbased highdimensional clustering and its application to microarray data. *Biometrics* **64**, 440-448.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *J. Amer. Statist. Assoc.* **105**, 713-726.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* **55**, 2183-2202.

Xu, Z., Chang, X., Xu, F. and Zhang, H. (2012). $L_{1/2}$ regularization: A thresholding representation theory and a fast solver. *IEEE Trans. Neural Netw. Learn. Syst.* **23**, 1013-1027.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541-2563.

Center of Data Science and Information Quality, Department of Information System and E-Business, School of Management, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

E-mail: xiangyuchang@gmail.com

Department of Statistics, University of California, Berkeley, Berkeley, California 94720, USA.

E-mail: shifwang@gmail.com

Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529, USA.

E-mail: rongjianli1985@gmail.com

Department of Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

E-mail: zbxu@mail.xjtu.edu.cn

(Received July 2015; accepted March 2017)