Statistica Sinica Preprint No: SS-2015-0261R3	
Title	Sparse k-Means with ℓ_{∞}/ℓ_0 Penalty for High-Dimensional Data Clustering
Manuscript ID	SS-2015-0261R3
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0261
Complete List of Authors	Rongjian Li Xiangyu Chang Yu Wang and Zongben Xu
Corresponding Author	Rongjian Li
E-mail	rli@cs.odu.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

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 new 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, Sparse k-means, Screening Property

1. Introduction

Clustering is an unsupervised technique for discovering hidden group structures from data sets. It partitions a whole sample set into different 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 et al., 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 real 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 the dimension reduction, such as principle components analysis (PCA) (Chang, 1983) or nonnegative matrix factorization (NMF) (Lee and Seung, 1999), before clustering algorithms are applied . However, existing evidences showed that these methods could not provide reasonable partition of the original data (Chang, 1983). Another idea is to perform the 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 of such idea is the high computation cost for training the model when the number of features is very large.

Recently, Witten and Tibshirani (2010) proposed a framework of sparse clustering which 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. In this paper, we call their method as 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 many synthetic data is good, a considerable portion of noise features is still kept in the final clustering result,

which is reported in (Witten and Tibshirani, 2010).

In this paper, we propose a new 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. However, the direct application of the ℓ_0 penalty on the sparse clustering framework (Witten and Tibshirani, 2010) will result in a solution that cannot be interpreted and theoretically analyzed explicitly. To address such challenges, we propose to jointly use both 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 theory. In order to justify the effectiveness of our proposed method on clustering field, we design multiple groups of experiments on synthetic data and real application data both. We show that ℓ_0-k-means exhibits much better noise feature detection capacity compared with ℓ_1-k-means.

Another important research topic in the high-dimensional statistics is analyzing the model behavior when the number of features (variables) grows with the sample size. In the literatures (Zhao and Yu, 2006; Wainwright, 2009), the authors proved the variable selection consistency property of Lasso. Negahban et al. (2012) developed a unified framework for analyzing error bounds of M-estimators with the decomposable regularizers, and Fan and Lv (2010) reviewed the techniques about the variable selection for penalized regression approaches. However, most of existing achievements can be categorized in the supervised learning field. The theoretical analysis for high-dimensional

data clustering method, which is an unsupervised learning method, is still limited (Pan and Shen, 2007; Witten and Tibshirani, 2010). To fill this blank, we discuss theoretical properties of ℓ_1 and ℓ_0-k-means in this paper. First, we verify that ℓ_1 and ℓ_0-k-means can be both interpreted from a thresholding perspective. Second, we further justify 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 a new one which includes the ℓ_0-k-means. We also give an efficient iterative algorithm to solve ℓ_0-k-means. We further 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. All the detailed proofs of theoretical results that are 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 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, $i = 1, \cdots, n$, are samples and columns $\mathbf{X}_j, j = 1, \ldots, p$ are features. It is well known that the 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,

$$\min_{\mathcal{C},\mu} \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mu_k), \tag{1.1}$$

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 formula of 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 could rewrite (1.1) in a form as follows:

$$\max_{\mathcal{C}} \sum_{j=1}^{p} \Big\{ \frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \Big\}, \tag{1.2}$$

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

$$a_j \triangleq \frac{1}{n} \sum_{i,i'}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j}, j = 1, \ldots, p, \tag{1.3}$$

then a_j is the jth component of BCSS, which can be considered as a function only with respect to the sample values of the jth feature and the partition $\mathcal{C}$. Note that we here used a_j to denote $a_j(\mathcal{C})$ for simplicity without causing any confusions. With the formulation (1.3), Witten and Tibshirani (2010) generalized the optimization problem with BCSS (1.2) as

$$\max_{\Theta(\mathcal{C}) \in D} \sum_{j=1}^{p} f_j(\mathbf{X}_j, \Theta(\mathcal{C})), \tag{1.4}$$

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*

$$\begin{aligned} \max_{\mathbf{w}, \Theta(\mathcal{C}) \in D} \quad & \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq 1, \ \|\mathbf{w}\|_1 \leq s, \ w_j \geq 0, \ j = 1, \ldots, p, \end{aligned} \tag{1.5}$$

where s is a tunning parameter, $\|\cdot\|_2$ is the ℓ_2-norm, $\|\cdot\|_1$ is the ℓ_1-norm, and $\mathbf{w} = (w_1, w_2, \ldots, w_p)^\top$ is a weight vector. Here, w_j can be interpreted as the contribution of the jth feature to the objective function (1.5). Moreover, they replaced $f_j(\mathbf{X}_j, \Theta(\mathcal{C}))$ by a_j defined as in (1.3), then (1.5) becomes the following ℓ_1-k-means model:

$$\begin{aligned} \max_{\mathcal{C},\mathbf{w}} \quad & \sum_{j=1}^{p} w_j \Big(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \Big) \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \le 1,\ \|\mathbf{w}\|_1 \le s,\ w_j \ge 0,\ \forall j = 1, \ldots, p. \end{aligned} \tag{1.6}$$

Although ℓ_1-k-means has been shown excellent performance on a sequence of experiments (Witten and Tibshirani, 2010), but there are still two disadvantages need to be considered. One is that it is inevitable to still keep amount of noise features (Wang et al., 2013). Witten and Tibshirani (2010) showed an example to describe this phenomenon: when 60 observations were generated from 3 clusters involving 50 relevant features and 150 noise features, ℓ_1-k-means kept all the noise features in the final clustering result. The other disadvantage is that neither the intuitive explanations on why it can select relevant features nor any theoretical guarantee about the property of ℓ_1-k-means were clearly supplied in literature. In this paper, we propose a new sparse k-means clustering framework to overcome these two drawbacks of ℓ_1-k-means.

2.2 ℓ_0-k-means

In literature, the ℓ_1 penalty is usually replaced by $\ell_q (0 \le q < 1)$ penalty for many sparse modeling problems when more sparsity is needed (Xu et al., 2012; Marjanovic and Solo, 2012; Wang et al., 2013). However, this substitution might be not such trivial and tractable for sparse clustering. For example, if we use the ℓ_0 penalty instead in (1.5),

this leads to the following optimization problem,

$$\begin{aligned} \max_{\mathbf{w},\Theta(\mathcal{C})} \quad & \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \le 1,\ \|\mathbf{w}\|_0 \le s, w_j \ge 0,\ j = 1, \ldots, p. \end{aligned} \tag{1.7}$$

Obviously, this model is not easy to analyze and solve since the objective function is no longer convex. To overcome this difficulty, we propose to jointly apply the ℓ_∞ and ℓ_0 penalty. In another words, we build the following new sparse clustering framework,

$$\begin{aligned} \max_{\mathbf{w},\Theta(\mathcal{C})\in D} \quad & \sum_{j=1}^{p} w_j f_j(\mathbf{X}_j, \Theta(\mathcal{C})) \\ \text{s.t.} \quad & \|\mathbf{w}\|_\infty \le 1,\ \|\mathbf{w}\|_0 \le s,\ w_j \ge 0,\ j = 1, \ldots, p, \end{aligned} \tag{1.8}$$

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 defined in (1.3). Thus, the final formulation of proposed ℓ_0-k-means is like the following,

$$\begin{aligned} \max_{\mathcal{C},\mathbf{w}} \quad & \sum_{j=1}^{p} w_j \Big(\frac{1}{n} \sum_{i=1}^{n} \sum_{i'=1}^{n} d_{ii'j} - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,i' \in C_k} d_{ii'j} \Big) \\ \text{s.t.} \quad & \|\mathbf{w}\|_\infty \le 1,\ \|\mathbf{w}\|_0 \le s,\ w_j \ge 0,\ j = 1, \ldots, p. \end{aligned} \tag{1.9}$$

We will show that this new ℓ_0-k-means method is not only tractable but also can be analyzed theoretically. Let us consider how to solve the ℓ_0-k-means (1.9). The difficulty mainly comes from the existence of two different types of variables: the partition variable $\mathcal{C} = \{C_1, \ldots, C_K\}$ which clustered 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 (1.9). To be specific, we solve $\mathbf{w}$ and

$\mathcal{C}$ alternatively by choosing one as the variable and fixing the other at the same time. Note that the iterative series is not guaranteed to converge to the global optimum, but the objective function will increase monotonically and achieve its maximal value. Since the sample can only be grouped in a finite number of ways and the optimal weights 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 optimum.

The detailed information about the solving procedure of ℓ_0-k-means is described in Algorithm 1.

In order to solve ℓ_0-k-means, we have to get the solution of optimization problem (1.10). Here, we provide the following Theorem 1 that could be used for the solving.

Theorem 1. *When the sequence $\{a_j\}_{j=1}^p$ defined by (1.3) is in a descending order, i.e., $a_i \geq a_j$ for any $i < j$, an optimal solution of (1.10) is given by*

$$w_j^* = \begin{cases} 1 & j \leq \lfloor s \rfloor \\ 0 & j > \lfloor s \rfloor \end{cases}, \tag{1.11}$$

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

Theorem 1 provides the solution of (1.10) with a closed-form. In another word, if $\{a_j\}_{j=1}^p$ is ordered, we can directly assign $w_j = 1$ for the components corresponding to the first $\lfloor s \rfloor$ elements of $\{a_j\}_{j=1}^p$ and $w_j = 0$ for the other elements. This procedure shows ℓ_0-k-means selects the relevant features by the gap information that is discussed in following Theorem 2.

We observe that the standard k-means costs $O(nKp)$ in time complexity and Step

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

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 $\mathbf{w}^{new}$:

$$\begin{aligned} \max_{\mathbf{w}} \quad & \mathbf{w}^\top \mathbf{a} \\ s.t. \quad & \|\mathbf{w}\|_\infty \le 1, \|\mathbf{w}\|_0 \le s, w_j \ge 0. \end{aligned} \tag{1.10}$$

4: Repeat step 2 and 3 until

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

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 which is same as the standard k-means. In fact, the condition $\lfloor s \rfloor \le nK$ is quite 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. Therefore, the ℓ_0-k-means is very efficient in implementation, which is also supported by experiments in the following section.

2.3 Theoretical Analysis

In this part, we analyze the theoretical properties of ℓ_1 and ℓ_0-k-means. We assume the data matrix is generated from a high-dimensional Gaussian mixture model. First, the ℓ_1 and ℓ_0-k-means are interpreted from a thresholding perspective. Then, we further prove that the solution of ℓ_1 and ℓ_0-k-means both have a screening consistent property under mild conditions. We also give theoretical comparison between these two models comprehensively.

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

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \phi_{ik} z_{ik}. \tag{1.12}$$

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

$$(\vec{v}_k)_j = \begin{cases} \mu_{kj} & j = 1, \ldots, p^* \\ 0 & j = p^* + 1, \ldots, p. \end{cases} \tag{1.13}$$

$\phi_{ik} \in \{0, 1\}$ is a binary random variable. $\mathbb{P}(\phi_{ik} = 1) = \pi_k$ and $\sum_{k=1}^{K} \phi_{ik} = 1$ for $k = 1, \ldots, K$. We assume each feature to be unitary with zero expectation, i.e. $\sum_k \pi_k \mu_k = 0$ and $\Sigma_{jj} = 1, j = 1, \ldots, p$. We further assume for each feature $j = 1, \ldots, p^*$, there exists at least two different 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. In another word, we assume the first p^* features are the relevant features. Denote $\mathcal{C}^* = \{C_1^*, \ldots, C_K^*\}$ to 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 (1.12) and (1.13),*

then

$$\mathbb{E}[a_j(\mathcal{C}^*)] = \begin{cases} K-1+c_j & 1 \le j \le p^*, \\ K-1 & \textit{otherwise}, \end{cases} \tag{1.14}$$

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

Theorem 2 reveals that there exists 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 becomes bigger and bigger when the K groups are distinguished more clearly on the jth feature.

Furthermore, we should mention that ℓ_1-k-means proposed in (Witten and Tibshirani, 2010) is in fact based on such gap to distinguish relevant features from noise features. Specifically, given an estimated partition $\widehat{\mathcal{C}}$, ℓ_1-k-means defined the optimal feature weight

$$\widehat{\mathbf{w}} = \frac{S(\mathbf{a}(\widehat{\mathcal{C}}), \Delta)}{\|S(\mathbf{a}(\widehat{\mathcal{C}}), \Delta)\|_2}, \tag{1.15}$$

where $S(\mathbf{a}, \Delta)_j = \max(a_j - \Delta, 0)$ is the well-known soft thresholding function (Donoho, 1995). From (1.15), we can see that any feature with $a_j < \Delta$ is identified as a noise feature, otherwise it is a relevant feature. Comparing with ℓ_1-k-means, Theorem 1 actually indicates that ℓ_0-k-means uses 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 will show their

feature selection capacity is different.

Let $\mathcal{C}$ be any partition of the n samples, and its BCSS for feature j be (1.3). By Lemma 1 in the supplementary materials, we know

$$a_j(\mathcal{C}) = -\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\right)^2 + \sum_{k=1}^{K}\left(\frac{1}{\sqrt{|C_k|}}\sum_{i\in C_k} x_{ij}\right)^2. \tag{1.16}$$

Now we omit the constant term and define the weighted BCSS as

$$\begin{aligned} F(\mathcal{C}, w) &\triangleq \sum_{j=1}^{p} w_j \bar{a}_j(\mathcal{C})^2 \\ &\triangleq \sum_{j=1}^{p} w_j \left(a_j(\mathcal{C}) + \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_{ij}\right)^2\right) \qquad (1.17) \\ &= \sum_{j=1}^{p} w_j \sum_{k=1}^{K}\left(\frac{1}{\sqrt{|C_k|}}\sum_{i\in C_k} x_{ij}\right)^2. \qquad (1.18) \end{aligned}$$

Our goal is to analyze the screening property of the following problem,

$$\begin{aligned} \max_{\mathcal{C},\mathbf{w}} \quad & F(\mathcal{C}, \mathbf{w}) \qquad (1.19) \\ s.t. \quad & \mathbf{w} \in \Omega, \end{aligned}$$

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

Definition 1. We claim the estimated weight $\widehat{\mathbf{w}}$ of (1.19) has the screening consistent property (SCP) if the following condition is satisfied

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

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

Now, we have the following theorems.

Theorem 3. *Denote* $(\widehat{\mathcal{C}}, \widehat{\mathbf{w}})$ *be the optimal solution of problem (1.19) where* $\Omega = \Omega_1 = \{\mathbf{w}| \, \|\mathbf{w}\|_1 \leq s, \|\mathbf{w}\|_2 \leq 1\}$. *Denote* $\sigma_1 = \min\limits_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$, $\sigma_2 = \max\limits_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$. *If* $p^{*2} \leq \frac{\sigma_1^4}{6400\sigma_2^3 \ln(K)}$ *and* $\ln(p) = o(n)$, *with* $\frac{\sum_{j=1}^{p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 - \frac{1}{2}\sigma_1 p^*}{\sqrt{\sum_{j=1}^{p^*} (\sum_{k=1}^K \pi_k \mu_{kj}^2 - \frac{1}{2}\sigma_1)^2}} \leq s \leq \frac{\sum_{j=1}^{p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2}{\sqrt{\sum_{j=1}^{p^*} (\sum_{k=1}^K \pi_k \mu_{kj}^2)^2}}$, *we have*

$$\mathbb{P}\left(\widehat{\mathbf{w}} \text{ has SCP}\right) \to 1, \text{as } n \to \infty \tag{1.20}$$

Theorem 4. *Denote* $(\widehat{\mathcal{C}}, \widehat{\mathbf{w}})$ *be the optimal solution of problem (1.19) where* $\Omega = \Omega_2 = \{\mathbf{w}| \, \|\mathbf{w}\|_0 \leq s, \|\mathbf{w}\|_\infty \leq 1\}$. *Denote* $\sigma_1 = \min\limits_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$, $\sigma_2 = \max\limits_{j=1,\ldots,p^*} \sum_{k=1}^K \pi_k \mu_{kj}^2 > 0$. *If* $p^{*2} \leq s^2 \leq \frac{\sigma_1^2}{192 \ln(K) \sigma_2}$ *and* $\ln(p) = o(n)$, *then*

$$\mathbb{P}\left(\widehat{\mathbf{w}} \text{ has SCP}\right) \to 1, \text{as } n \to \infty \tag{1.21}$$

Theorem 3 and 4 indicate that ℓ_1 and ℓ_0-k-means both have the SCP if p^* is small enough and $\ln p = o(n)$. Note that the condition $\ln p = o(n)$ means the number of features grows slower than an exponential order with respect to the sample size. Such property was considered to be optimal for regularized regression approaches about 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 different. We will compare them in the next section.

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 a real application data from the biology field which is Allen Developing Mouse Brain Atlas.

ℓ_0-k-means involves a tuning parameter s which controls the number of features selected. Witten and Tibshirani (2010) proposed a strategy to select the tunning pa-

rameter s based on the gap statistic (Tibshirani et al., 2001). In this paper, we follow their strategy for the proposed ℓ_0-k-means as well. We consider two different criteria to obtain comprehensive comparisons. The first criterion is the *Classification Error Rate* (CER) (Witten and Tibshirani, 2010; Chipman and Tibshirani, 2006), which is 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 a same group with respect to partition $\mathcal{C}$. The second criterion is F_1-score, which measures the feature selection accuracy. If we define precision to be

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

and recall to be

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

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

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

3.1 Evaluation on Synthetic Data

In this part, four numerical experiments are conducted. The first experiment is to verify that the gap statistic succeeds in selecting an appropriate tuning parameter for ℓ_0-k-means. The second and the third experiment are 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 explore the performance of those algorithms for non-Gaussian distributions.

Experiment 1: We construct 6 clusters and each cluster contains 20 samples with 2000 features, which brings a data matrix $\mathbf{X}_{120\times 2000}$. Among these 2000 features, we assume only first 200 features are relevant features. For the kth cluster, relevant features are sampled from a Gaussian distribution $\mathcal{N}(0.5\cdot k, 1)$ and noise features are sampled from $\mathcal{N}(0,1)$ separately. The data matrix is normalized to have column-wise zero mean before any algorithm is applied. We repeat 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. All the results are shown in Figure 1.

Figure 1 summarizes all the results of ℓ_0-k-means compared with standard k-means. From the left subfigure, 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 is 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. We can observe that the values for relevant features are approximately close to 1 while those for noise features are close to 0. This shows that the usage of gap statistic for ℓ_0-k-means can help the selection of relevant features and 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 is generated from a Gaussian mixture model with independent features. We further assume each element x_{ij} in data

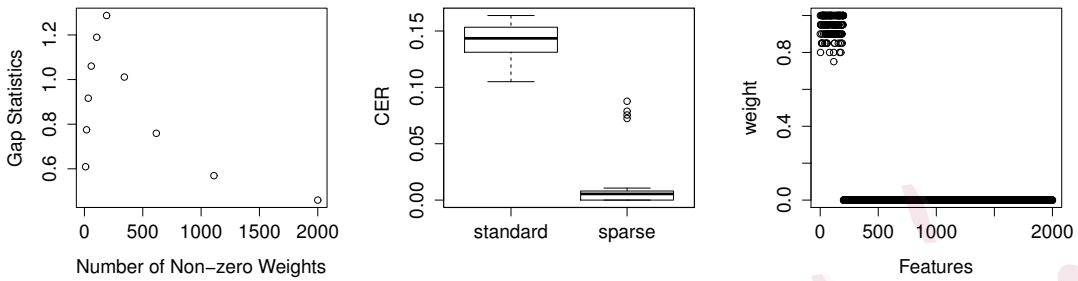

Figure 1: Overview of ℓ_0-k-means.

matrix is drawn from the distribution $\mathcal{N}(\mu_{ij}, \sigma_j^2)$ independently, where

$$\mu_{ij} = \begin{cases} a_j\mu & \text{if } i \in C_1, j \leq 50 \\ -a_j\mu & \text{if } i \in C_2, j \leq 50 \\ 0 & \text{if } i \in C_3, \text{ or } j > 50 \end{cases}, \tag{1.22}$$

where a_j is chosen randomly from $[0.75, 1.25]$ for each $j = 1, \ldots, 50$ and σ_j is chosen randomly from $[0.75, 1.25]$ for $j = 1, \ldots, p$. In other words, the first 50 features are relevant features while the rest ones are noise features. There are 3 clusters and each cluster contains 50 samples. $\mu = 0.6, 0.7$ and $p = 200, 500, 1000$. Each parameter setting repeats 50 times. The results are reported in Figure 2 and Figure 3.

In Figure 2, ℓ_0-k-means has 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. It can be seen that ℓ_0-k-means, compared to other algorithms, has F_1-score close to 1 with a very small deviation. This might explain why ℓ_0-k-means tends 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 is generated from a Gaussian mixture model with correlated features. Suppose

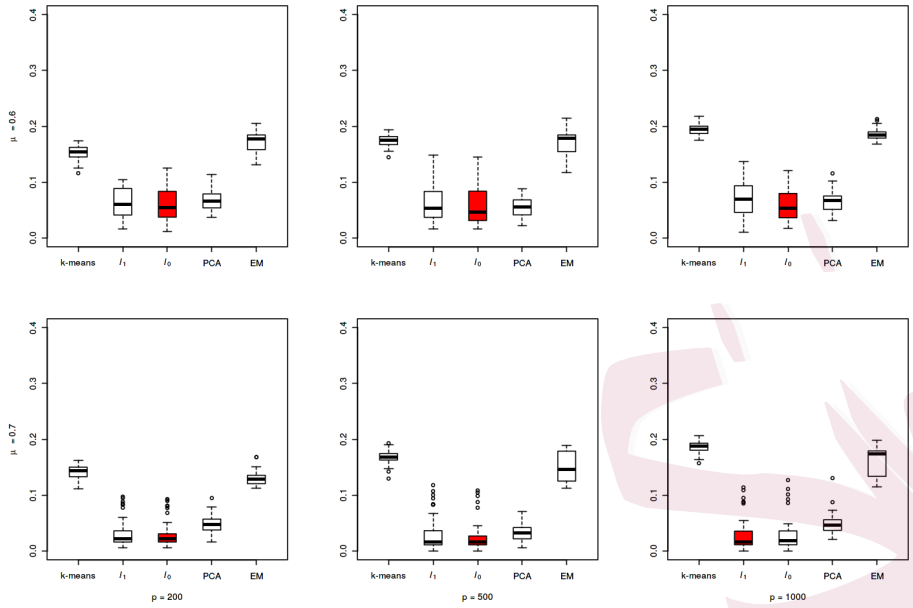

Figure 2: *CER Boxplot for Experiment 2*

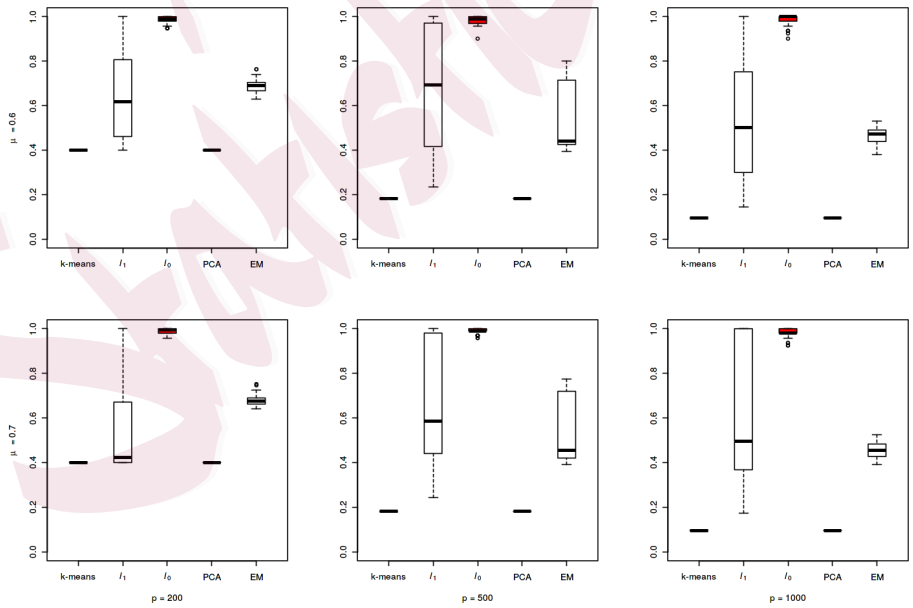

Figure 3: *F_1-score Boxplot for Experiment 2*

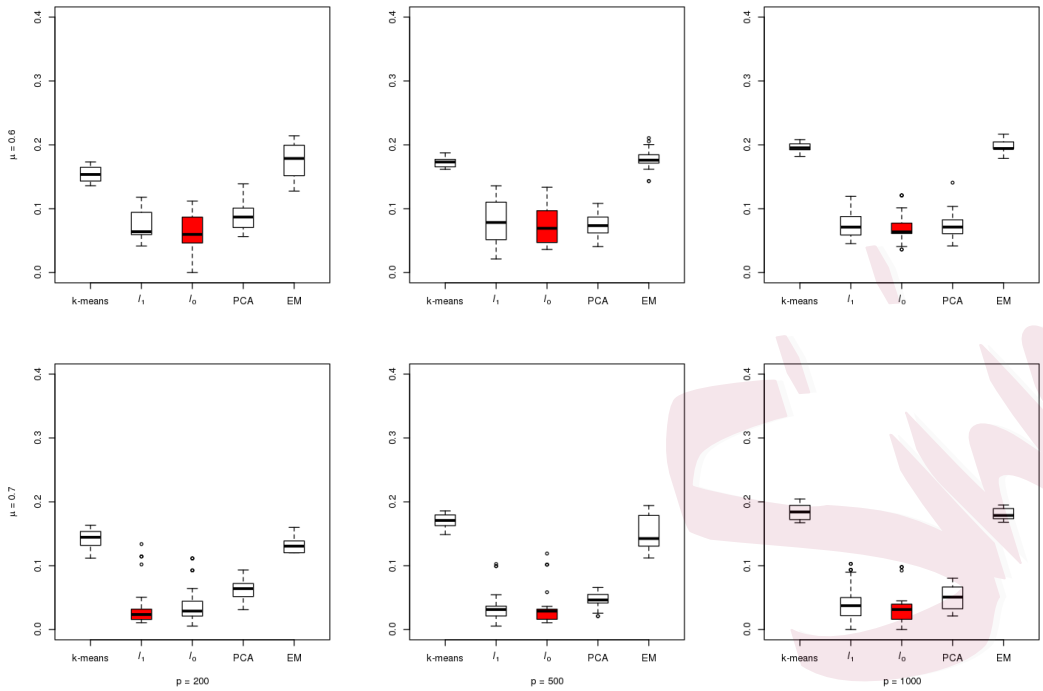

Figure 4: *CER Boxplot for Experiment 3*

each sample $\mathbf{x}_i$ is drawn from a distribution $\mathcal{N}(\mu, \Sigma)$, where the elements Σ_{ij} of Σ is set to be $\Sigma_{ij} = 0.1^{|i-j|}$.

In Figure 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 other algorithms.

Experiment 4: In this experiment, we extend the Gaussian mixture model to non-Gaussian cases. Experiment settings are identical to Experiment 2 except we use standard log normal distribution $f(x) = k \cdot \mu + a \cdot \exp(\mathcal{N}(0, 1))$ and standard Poisson distribution $f(x) = k \cdot \mu + \text{Poisson}(1)$, where a is chosen randomly from $[0.75, 1, 25]$ and $k = 1, \ldots, K$. Set $\mu = 2, 3$ for the log normal distribution and $\mu = 1, 1.5$ for the Poisson

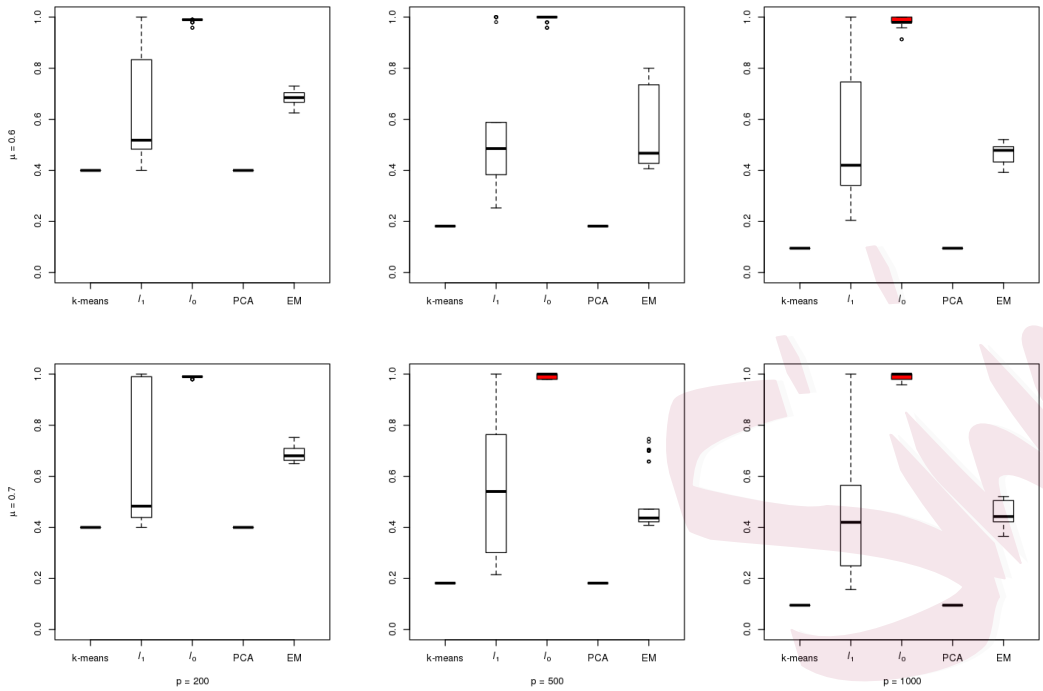

Figure 5: *F_1-score Boxplot for Experiment 3*

distribution. The results are shown in Figure 6 to Figure 9. Surprisingly, as shown in the Figure 7 and Figure 9, ℓ_0-k-means still achieves the best feature selection accuracy.

3.2 Evaluation on Allen Developing Mouse Brain Atlas

In this section, we compare our proposed method with other existing methods on a real application data which is Allen Developing Mouse Brain Atlas data (Lein et al., 2006; 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

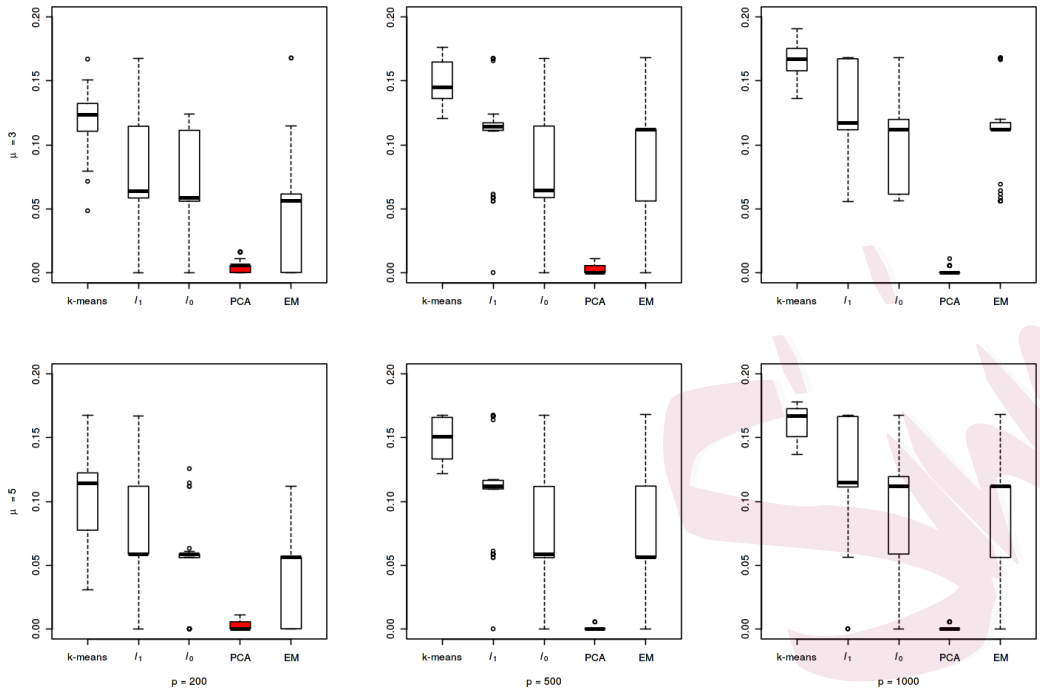

Figure 6: *CER Boxplot for Experiment 4 log normal case*

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 as 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 apply 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) respectively to the 7 data matrices. The detailed results including CER values and the feature selection performance are shown

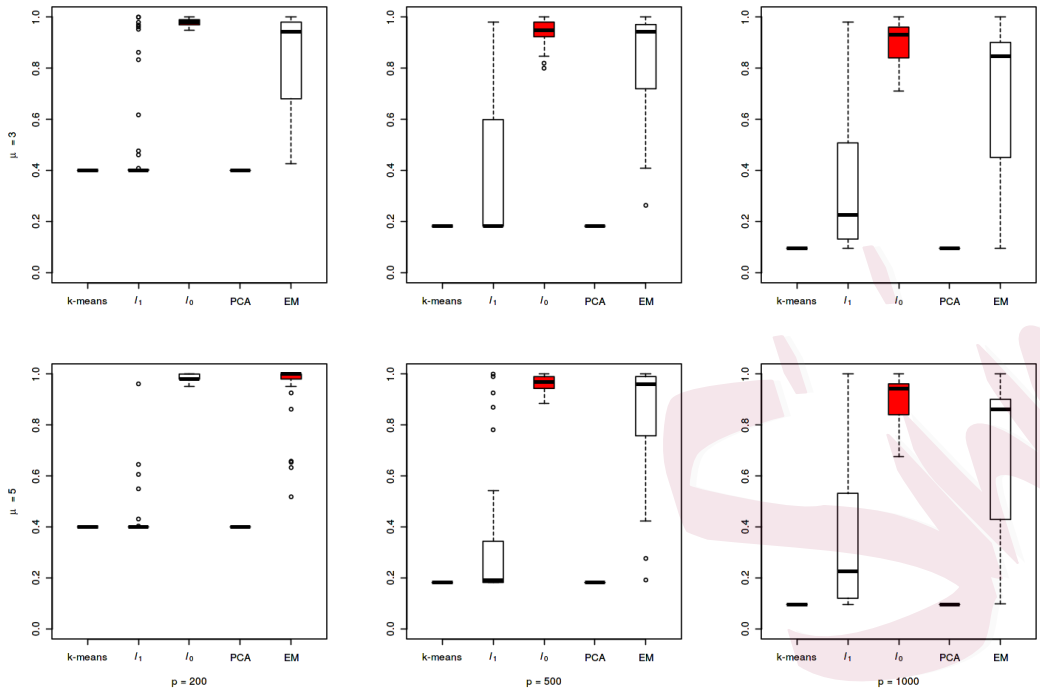

Figure 7: *F_1-score Boxplot for Experiment 4 log normal case*

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	1724	1724	1724	1724	1724	1724	1724
Number of voxels	7122	13194	12148	12045	21845	24180	28023
Number of regions	20	20	20	20	20	19	20

in Table 2 and 3. From Tables 2, we can see that the ℓ_0-k-means in most cases outperforms other competitors. Besides the low CER values while using minimal the number of features (i.e., nonzero weights $\mathbf{w}$), another advantage of ℓ_0-k-means is its *interpretability*, which means that ℓ_0-k-means can achieve the almost smallest CER value even it

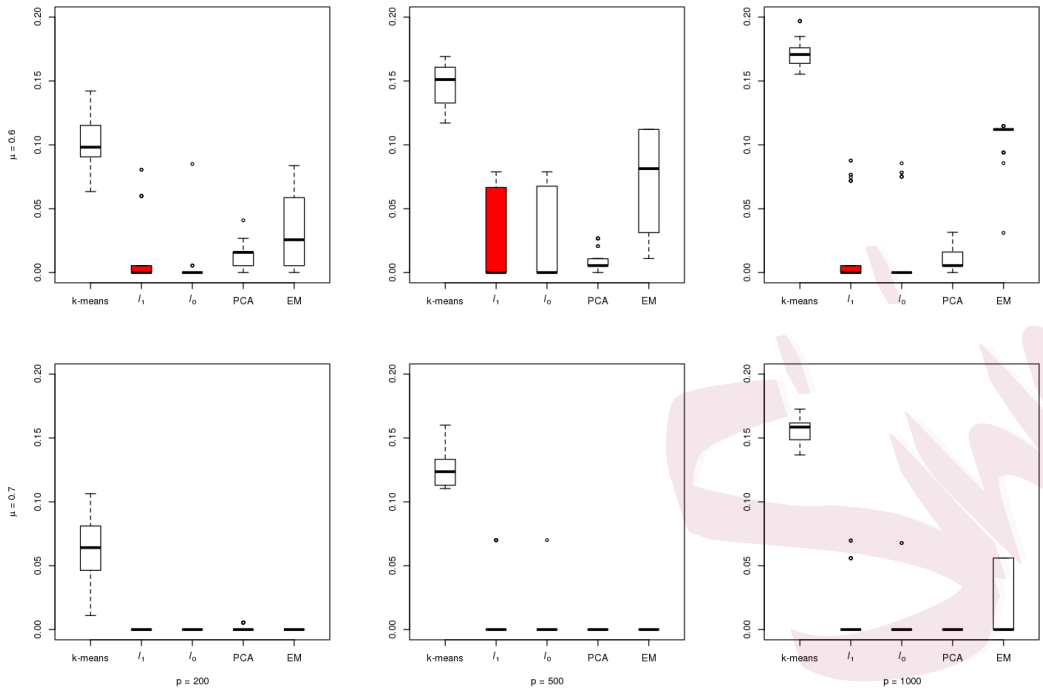

Figure 8: *CER Boxplot for Experiment 4 Poisson case*

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**

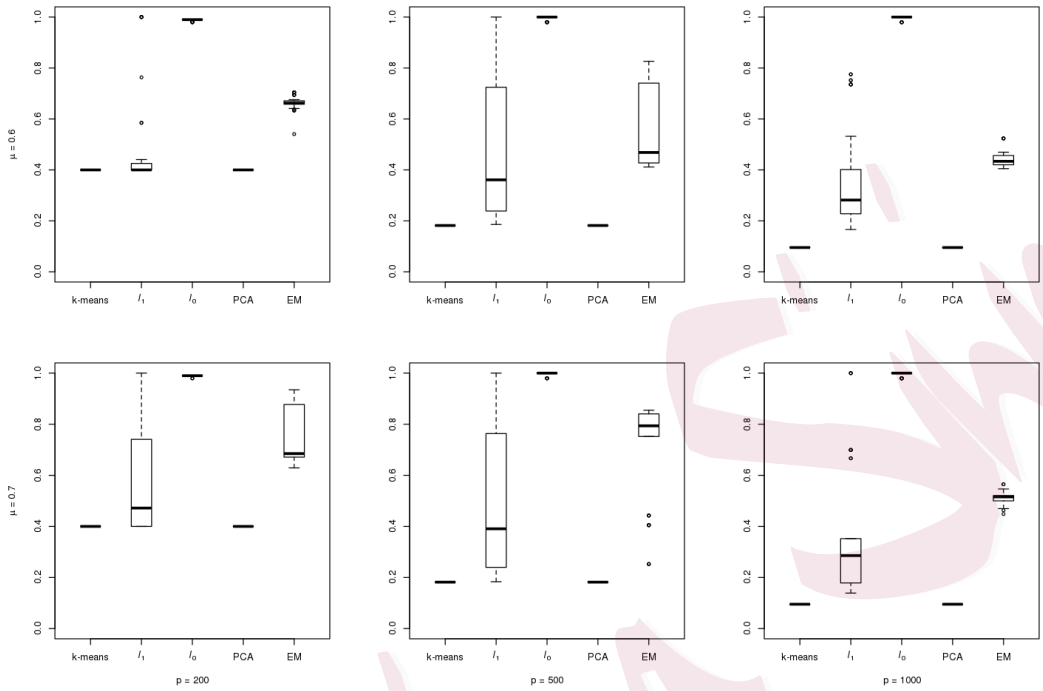

Figure 9: *F_1-score Boxplot for Experiment 4 Poisson case*

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

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

Ages	E11.5	E13.5	E15.5	E18.5	P4	P14	P28
k-means	1723	1724	1724	1724	1720	1724	1724
ℓ_1-k-means	717	672	659	642	446	224	1724
ℓ_0-k-means	100	660	100	1600	199	322	1068
PCA-k-means	1723	1724	1724	1724	1720	1724	1724
EM	1723	1724	1724	1724	1720	1724	1724

employs fewest features. The reason may be that ℓ_0-k-means can eliminate more noise features compared with other methods. For instants, we consider the postnatal stages P14 because the differentiation of gene functions is much more discriminative at this postnatal stage. We observe that there is few "noisy" genes which have been eliminated by ℓ_0-k-means and included by ℓ_1-k-means. For example, we find a noisy gene of name 'Scn4b' which is detected by our ℓ_0-k-means method. This gene is highly related with the protein composition of sodium channel beta subunits (Medeiros-Domingo et al., 2007). In another word, it is strongly bonded with the electrical signal transmission activities in most of types of cells. Therefore, it is reasonable to consider the feature corresponding to this gene as a noise feature, since its function is uniformly supportive in the whole brain and using it to distinguish different regions might not be effective. Therefore, detecting the feature as a noise feature by ℓ_1-k-means is consistent with the prior knowledge about

genes listed in the database of Allen Institute*.

Overall, the experiments demonstrate that ℓ_0-k-means exhibits an outperforming capacity on noise feature detection.

4. Conclusion and future work

In this paper, we focus on designing a novel efficient clustering algorithm for high dimensional datasets. Inspired by the literature in sparse clustering, we allow algorithms to optimize weights of individual features to combine clustering procedure with feature selection. Moreover, we proposed a new optimization formulation with ℓ_∞/ℓ_0 penalty, which we call ℓ_0-k-means. ℓ_0-k-means can be efficiently solved by Algorithm 1. Although theoretical analysis reveals that both ℓ_0-k-means and ℓ_1-k-means have screening consistency under appropriate conditions for Gaussian mixture model, empirical experiments suggests that ℓ_0-k-means outperforms ℓ_1-k-means in feature selection in terms of F_1-score. Extensive experiments are carried out to compare with some other well-known clustering methods.

In the future, we might carry out our work in the following directions. First of all, we will 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 will also extend the current research by applying on other high-dimensional data clustering model, for instance, the penalized model-based clustering (Pan and Shen, 2007). In addition, the proposed method could be applied on other real life datasets to show its effectiveness.

Supplementary Materials

We provide the detailed proofs of the proposed theorems in the online supplementary

*http://www.genecards.org/

material.

Acknowledgement

We would like to thank the review team-the editors and the two anonymous reviewers, for your careful work and constructive comments, which greatly help us improving the paper quality. 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. (2006). High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete Comput. Geom.* **35**, 617-652.

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

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

REFERENCES

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.

REFERENCES

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. Series 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, USA

E-mail: shifwang@gmail.com

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

REFERENCES

E-mail: rli@cs.odu.edu

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

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