Electronic Journal of Statistics
Vol. 6 (2012) 148–167
ISSN: 1935-7524
DOI: 10.1214/12-EJS668

Regularized k-means clustering of high-dimensional data and its asymptotic consistency

Wei Sun* **and Junhui Wang**

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
e-mail: wsun8@uic.edu; jwang@math.uic.edu

Yixin Fang

Division of Biostatistics, School of Medicine
New York University
New York, NY 10016
e-mail: Yixin.Fang@nyumc.org

Abstract: K-means clustering is a widely used tool for cluster analysis due to its conceptual simplicity and computational efficiency. However, its performance can be distorted when clustering high-dimensional data where the number of variables becomes relatively large and many of them may contain no information about the clustering structure. This article proposes a high-dimensional cluster analysis method via regularized k-means clustering, which can simultaneously cluster similar observations and eliminate redundant variables. The key idea is to formulate the k-means clustering in a form of regularization, with an adaptive group lasso penalty term on cluster centers. In order to optimally balance the trade-off between the clustering model fitting and sparsity, a selection criterion based on clustering stability is developed. The asymptotic estimation and selection consistency of the regularized k-means clustering with diverging dimension is established. The effectiveness of the regularized k-means clustering is also demonstrated through a variety of numerical experiments as well as applications to two gene microarray examples. The regularized clustering framework can also be extended to the general model-based clustering.

AMS 2000 subject classifications: Primary 62H30.
Keywords and phrases: K-means, diverging dimension, lasso, selection consistency, variable selection, stability.

Received April 2011.

Contents

*The authors would like to thank the editor, the associate editor and the referees for their insightful comments and suggestions

1. Introduction

Cluster analysis is to assign observations into a number of clusters such that observations in the same cluster are similar to each other. The similarity is often quantified by some distance measures, such as the Euclidean distance [15] and correlation [2]. To optimize the similarity measures, various clustering algorithms are developed. Among others, k-means clustering is one of the most popular clustering algorithms, which aims at minimizing the within-cluster dissimilarity measured by the Euclidean distance. While the k-means clustering is conceptually simple and computationally efficient, its performance can be severely deteriorated when clustering high-dimensional data where the number of variables becomes large and many of them may contain no information about the clustering structure. Furthermore, the interpretability of the k-means clustering can be impeded as it usually includes all the variables and produces complicated clustering models. To overcome these difficulties in clustering high-dimensional data, a more appropriate clustering algorithm that can simultaneously perform cluster analysis and select informative variables is in demand.

In statistical literature, two major kinds of variable selection techniques are developed in the context of high-dimensional data analysis. The first kind is to pre-screen the redundant variables by conducting a multiple testing procedure and controlling certain error rates, such as [7] and the reference therein. The second kind is the shrinkage method, which penalizes the model fitting with various types of regularization terms that encourage model sparsity, such as the LASSO regression in [20]. Although variable selection for regression has been extensively studied, analogous result for clustering is limited, such as [19, 16, 26, 24, 12, 10]. Focusing on the k-means and hierarchical clustering, [25] proposed a general sparse clustering framework using a similar idea as nonnegative garrote [4], however the asymptotic consistency was not discussed in their framework.

In this article, we propose a regularized k-means clustering, which can perform cluster analysis and variable selection at the same time. The key idea is to formulate k-means clustering in a form of regularization, with an adaptive group lasso penalty term on cluster centers. Note that all cluster centers share the same set of variables, so the group lasso penalty term is employed to select

the variables in a group fashion; i.e., a variable is redundant if it is not used in any cluster center. The regularized k-means clustering framework can also be extended to the model-based clustering, where the EM algorithm is employed to minimize the regularized negative log-likelihood function. To optimally balance the trade-off between model fitting and sparsity, a model selection criterion is developed based on the clustering stability in [3, 23]. The key idea is that if multiple samples are available from the same distribution, a good clustering algorithm should yield clustering assignments of observations that do not vary much from one sample to another. An efficient estimation scheme based on bootstrap is proposed to accurately estimate the clustering stability in high-dimensional clustering. Furthermore, the asymptotic estimation and selection consistency of the proposed regularized k-means clustering with diverging dimension is established. Whereas the selection consistency in regression has been obtained in [9, 29], analogous results in the context of cluster analysis seem rare. The effectiveness of the proposed algorithms is also demonstrated in a variety of simulated examples as well as applications to two gene microarray examples.

The rest of the paper is organized as follows. Section 2 reviews the standard k-means clustering. Section 3 presents the proposed regularized k-means clustering as well as its efficient implementation. Section 4 introduces the stability-based model selection criterion for tuning the regularized k-means clustering. Asymptotic estimation and selection consistency is established in section 5. Extension to the regularized model-based clustering is provided in section 6, followed by simulation studies in section 7 and two real gene examples in section 8. A brief discussion is given in section 9. Technical details are provided in Appendix.

2. Clustering analysis and k-means clustering

In the k-means clustering, assume that n data points $X_1, \ldots, X_n$ are available with $X_i = (X_{i1}, \ldots, X_{ip})^T$, and the number of clusters is pre-specified as K. The K clusters are denoted by $\mathcal{A}_1, \ldots, \mathcal{A}_K$ with centers $C_1, \ldots, C_K$, where $C_k = (C_{k1}, \ldots, C_{kp})^T$. The k-means clustering then attempts to solve

$$\min_{\mathcal{A}_k, C_k} \sum_{k=1}^{K} \sum_{X_i \in \mathcal{A}_k} \|X_i - C_k\|^2, \tag{1}$$

where $\|\cdot\|$ is the standard Euclidean norm.

Note that the global minimization in (1) is NP-hard and requires integer programming due to the discrete feature of $\mathcal{A}_k$. As a remedy, an iterative scheme [14] is often employed to approximate the solution of (1), which updates $\mathcal{A}_k$ and C_k separately at each iteration pretending the other one is fixed. Specifically, at t-th iteration, for the fixed K centers $C_1^{(t-1)}, \ldots, C_K^{(t-1)}$, $\mathcal{A}_k^{(t)}$ is updated by assigning each observation X_i to the closest cluster; and then for the fixed $\mathcal{A}_k^{(t)}$, $C_k^{(t)} = |\mathcal{A}_k^{(t)}|^{-1} \sum_{X_i \in \mathcal{A}_k^{(t)}} X_i$, where $|\mathcal{A}_k^{(t)}|$ is the cardinality of $\mathcal{A}_k^{(t)}$.

Although the k-means clustering has been reported successful in many real applications, its performance can be less effective in high-dimensional cluster

analysis. [13] pointed out that when the sample size is fixed and the dimension diverges, the distances among observations tend to be deterministic. In specific, the observations from the same cluster tend to lie symmetrically at the vertices of a regular simplex, and the distance between observations from different clusters is determined by the cluster difference relative to the data dimension. Consequently, if the cluster difference is relatively small compared with the diverging data dimension, the k-means clustering based on the Euclidean distance will operate in a degenerate fashion, assigning all the observations to the same cluster. In addition, the k-means clustering tends to include all the variables no matter if the variable contains information about the clustering structure or not. This is undesirable in high dimensional cluster analysis, where the clustering structure often lies in a low dimensional subspace and the majority of the variables are redundant in capturing the structure.

3. Regularized k-means clustering

This section proposes the regularized k-means clustering for high dimensional cluster analysis, which allows simultaneous clustering model fitting and variable selection.

The key idea of the regularized k-means clustering is to extend the k-means clustering in (1) by adding an adaptive group lasso penalty term on cluster centers. Specifically, the regularized k-means clustering is formulated as

$$\min_{\mathcal{A}_k, C_k} \frac{1}{n} \sum_{k=1}^{K} \sum_{X_i \in \mathcal{A}_k} \|X_i - C_k\|^2 + \sum_{j=1}^{p} J(C_{(j)}), \tag{2}$$

where the training data $X_1, \ldots, X_n$ are centralized so that the mean of each variable is zero. In (2), the first term is equivalent to the k-means clustering, which measures the within-cluster distance from each observation to its corresponding cluster center, the second term $J(C_{(j)})$ is a regularization term on each variable, where $C_{(j)} = (C_{1j}, \ldots, C_{Kj})^T$ and C_{kj} is the j-th element of C_k. Particularly, the regularization term $J(C_{(j)})$ can be group LASSO penalty $\lambda\|C_{(j)}\|$ [27], or adaptive group LASSO penalty $J(C_{(j)}) = \lambda_j\|C_{(j)}\|$ [22], where λ and $\lambda_j, j = 1, \ldots, p$ are tuning parameters that control the balance between the clustering model fitting and sparsity. Whereas the group LASSO penalty uses the same λ for all dimensions and may ignore the relative importance of each dimension [28], the adaptive group LASSO penalty associates each dimension with a different λ_j so that the relative importance of each dimension can be incorporated. For illustration, we set $J(C_{(j)}) = \lambda_j\|C_{(j)}\|$ in this article and note that it can be generalized to other types of regularization terms such as the group LASSO penalty and the L_∞-norm penalty [24].

To solve the optimization in (2), we adopt a similar iterative scheme as in solving the k-means clustering. That is, we update $\mathcal{A}_k$ and C_k separately at each iteration pretending the other one is fixed. When C_k is fixed, $\mathcal{A}_k$ is updated by assigning each observation X_i to the closest cluster. When $\mathcal{A}_k$ is fixed, the

following Lemma 3.1 suggests that C_k can be solved in a componentwise fashion, which can substantially facilitate the computation in high-dimensional cluster analysis.

Lemma 3.1.

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{X_i\in\mathcal{A}_k}\|X_i - C_k\|^2 + \sum_{j=1}^{p} J(C_{(j)}) = \sum_{j=1}^{p}\Big(\frac{1}{n}(X_{(j)} - LC_{(j)})^T(X_{(j)} - LC_{(j)}) + J(C_{(j)})\Big),$$

where $X_{(j)} = (X_{1j}, \ldots, X_{nj})^T$ *is the* j*-th variable across all sample points,* L *is a cluster assignment matrix with* $L_{ik} = \mathbf{1}(X_i \in \mathcal{A}_k)$*;* $i = 1, \ldots, n$*,* $k = 1, \ldots, K$*, and* $\mathbf{1}(\cdot)$ *is an indicator function.*

Lemma 3.1 follows immediately from the following equality,

$$\begin{aligned}\sum_{k=1}^{K}\sum_{X_i\in\mathcal{A}_k}\|X_i - C_k\|^2 &= \sum_{i=1}^{n}\sum_{X_i\in\mathcal{A}_k}\sum_{j=1}^{p}(X_{ij} - C_{kj})^2 \\ &= \sum_{j=1}^{p}(X_{(j)} - LC_{(j)})^T(X_{(j)} - LC_{(j)}).\end{aligned}$$

A direct consequence of Lemma 3.1 is that when L is fixed, solving (2) can be simplified to

$$\min_{C_{(j)}} \frac{1}{n}(X_{(j)} - LC_{(j)})^T(X_{(j)} - LC_{(j)}) + J(C_{(j)}) \tag{3}$$

for each individual variable, where $J(C_{(j)}) = \lambda_j\|C_{(j)}\|$ with $\lambda_j = \lambda\|\widetilde{C}_{(j)}\|^{-1}$, and $\widetilde{C}_{(1)}, \ldots, \widetilde{C}_{(p)}$ are the estimated cluster centers from the standard k-means clustering.

The details of the proposed regularized k-means clustering are as follows.

Algorithm 1 (Regularized k-means clustering).

Step 1. Initialize centers $C_1^{(0)}, \ldots, C_K^{(0)}$ by the standard k-means clustering.

Step 2. Until the termination condition is met, repeat

(a). Given $C_1^{(t-1)}, \ldots, C_K^{(t-1)}$, find the cluster assignment matrix $L^{(t)}$.

(b). Given $L^{(t)}$, update $C^{(t)}$ by minimizing (3) for each j.

As computational remarks, to overcome the sensitivity to the initialization in Step 1 the standard k-means clustering is randomly started multiple times and the one with smallest within-cluster distance is selected as the initialization. In Step 2 the iteration stops when $L^{(t)}$ does not change any more. Based on our limited numerical experience, the algorithm stops often within no more than five iterations.

4. Selection of tuning parameters

In the proposed regularized k-means clustering formulation, two tuning parameters, K and λ, need to be appropriately determined so that the clustering performance can be optimized. In this section, the tuning parameters are selected through a selection criterion based on clustering stability.

The key idea of clustering stability is that if we repeatedly draw samples from the same population and apply the regularized clustering algorithm, a good clustering algorithm should produce clustering assignments that are similar from one sample to another. In the proposed regularized k-means clustering, different values of K and λ define different clustering algorithms, therefore we select the values of K and λ such that the resulting clustering algorithm has the maximal clustering stability.

Denote that $Z = \{X_1, \ldots, X_n\}$ is a random sample of size n from some unknown distribution $F(x)$ with $x \in R^p$. Following [23], we define clustering assignment $\psi(x)$ to be a mapping: $R^p \to \{1, \ldots, K\}$, and the regularized k-means clustering $\Psi(\cdot; K, \lambda)$ generates a clustering assignment ψ when applied to a sample Z. The clustering distance between any two clustering assignments $\psi_1(x)$ and $\psi_2(x)$ is defined as

$$d(\psi_1, \psi_2) = P\Big(\{\psi_1(X) = \psi_1(Y)\} \wedge \{\psi_2(X) = \psi_2(Y)\}\Big), \tag{4}$$

where X and Y are independently sampled from F, and $A \wedge B = (A \setminus B) \cup (B \setminus A)$. Clearly, the distance between ψ_1 and ψ_2 measures the probability of their disagreement. The clustering instability of regularized k-means clustering $\Psi(\cdot; K, \lambda)$ is then

EXPECTED DISTANCE

$$S(\Psi, K, \lambda, n) = E(d\{\Psi(Z_1; K, \lambda), \Psi(Z_2; K, \lambda)\}), \tag{5}$$

where $\Psi(Z_1; K, \lambda)$ and $\Psi(Z_2; K, \lambda)$ are clustering assignments obtained by applying $\Psi(\cdot; K, \lambda)$ to two independent samples Z_1 and Z_2 respectively.

To accurately estimate $S(\Psi, K, \lambda, n)$, we propose the bootstrap resampling scheme. Consider the candidate algorithms $\{\Psi(\cdot, K, \lambda) : K = 2, \ldots, K.max; \lambda \geq 0\}$, where $K.max$ specifies the largest possible number of clusters, and $K = 1$ is excluded as it assigns all observations into the same cluster and thus provides little structural information of the data. Given n observations $(X_1, \ldots, X_n)$, three independent bootstrap samples of the same size n, $Z_1^{*b}, Z_2^{*b}, Z_3^{*b}$, are generated, where $b = 1, \ldots, B$ denotes the b-th replication. Two clustering assignments, $\Psi(Z_1^{*b}; K, \lambda)$ and $\Psi(Z_2^{*b}; K, \lambda)$ are constructed based on Z_1^{*b} and Z_2^{*b} respectively, and $S(\Psi, K, \lambda, n)$ is estimated as the distance between $\Psi(Z_1^{*b}; K, \lambda)$ and $\Psi(Z_2^{*b}; K, \lambda)$ on Z_3^{*b},

$$\begin{aligned} &\widehat{S}^{*b}(\Psi, K, \lambda, n) \\ = \ &\binom{n}{2}^{-1} \Big| (i,j)_{i<j} : I(\widehat{\psi}_1^{*b}(X_i^{(3)}) = \widehat{\psi}_1^{*b}(X_j^{(3)})) \neq I(\widehat{\psi}_2^{*b}(X_i^{(3)}) = \widehat{\psi}_2^{*b}(X_j^{(3)})) \Big|, \end{aligned}$$

where $\widehat{\psi}_1^{*b} = \Psi(Z_1^{*b}; K, \lambda)$ and $\widehat{\psi}_2^{*b} = \Psi(Z_2^{*b}; K, \lambda)$, $X_i^{(3)}$ and $X_j^{(3)}$ are elements in sample Z_3^{*b}, and $|A|$ is the cardinality of set A. Then the optimal K and λ can be estimated by the following voting scheme. For each λ, $\widehat{K}_\lambda = \text{mode}\{\widehat{K}_\lambda^{*1}, \ldots, \widehat{K}_\lambda^{*B}\}$, where $\widehat{K}_\lambda^{*b} = \text{argmin}_{2 \le K \le K.max} \widehat{S}^{*b}(\Psi, K, \lambda, n)$, then the optimal K is estimated as $\widehat{K} = \text{mode}\{\widehat{K}_\lambda\}$. Given the estimated $\widehat{K}$, the optimal λ is estimated as $\widehat{\lambda} = \text{mode}\{\widehat{\lambda}^{*1}, \ldots, \widehat{\lambda}^{*B}\}$, where $\widehat{\lambda}^{*b} = \text{argmin}_\lambda \widehat{S}^{*b}(\Psi, \widehat{K}, \lambda, n)$.

5. Consistency of regularized k-means clustering

We now present the asymptotic estimation and selection consistency of the proposed regularized k-means clustering with diverging dimension. The estimation consistency assures that the estimated cluster centers converge almost surely to the true cluster centers based on population, and the selection consistency shows that the uninformative variables are eliminated from the estimated cluster centers with probability tending to one.

Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution P, and denote P_n as the associated empirical measure. Regarding (2) as a function of cluster centers and the empirical measure P_n, the regularized k-means clustering is to minimize

$$W(C, P_n) = \int \min_{C_k \in C} \|x - C_k\|^2 P_n(dx) + \sum_{j=1}^{p} J(C_{(j)}) \tag{6}$$

over $C = (C_1, \ldots, C_K)^T$. Denote $\widehat{C} = (\widehat{C}_1, \ldots, \widehat{C}_K)^T$ as the estimated cluster centers by solving (6), $\bar{C} = (\bar{C}_1, \ldots, \bar{C}_K)^T$ as the true cluster centers which minimizes

$$W(C, P) = \int \min_{C_k \in C} \|x - C_k\|^2 P(dx),$$

and $\widehat{L}$, $\bar{L}$ as the cluster assignment matrices of $X_1, \ldots, X_n$ based on $\widehat{C}$ and $\bar{C}$ respectively.

Theorem 1. *Under Assumptions (i) – (vi) in the Appendix, if* $n^{1/2}\lambda p \to 0$ *and* $n^{-2}\lambda^{-2} p \to 0$ *as* $n \to \infty$, *then* $\widehat{C} \to \bar{C}$ *almost surely and* $\|\widehat{C} - \bar{C}\| = O_p(n^{1/2}\lambda p^{-1})$.

Theorem 1 shows that the regularized k-means clustering with a properly selected λ attains similar asymptotic estimation consistency as the standard k-means clustering in [17, 18]. Note that the dimension p is allowed to diverge to infinity at an order of $o(\min(n^2\lambda^2, n^{-1/2}\lambda^{-1}))$. In specific, if $p = O(n^a)$ with $0 < a < 1/3$, setting $\lambda = O(n^{-(a+3)/4})$ satisfies the order conditions. These conditions have also been used in [9] for establishing the asymptotic consistency of high-dimensional regularized regression.

Next we establish the asymptotic selection consistency of the regularized k-means clustering, which is desirable in high-dimensional cluster analysis where many variables are redundant and contain no information about the clustering

structure. Without loss of generality, we assume that only the first $p_0 < p$ variables are informative in that $\|\bar{C}_{(j)}\| \neq 0$ for $j \leq p_0$ and $\|\bar{C}_{(j)}\| = 0$ for $j > p_0$. The informative variable set is denoted as $\mathcal{A} = \{1, \ldots, p_0\}$ and the uninformative variable set is then $\mathcal{A}^c = \{p_0 + 1, \ldots, p\}$.

Theorem 2. *Under Assumptions* $(i) - (vii)$ *in the Appendix, if* $n^{1/2}\lambda p \to 0$ *and* $n^{-2}\lambda^{-2}p \to 0$ *as* $n \to \infty$*, then* $P(\|\widehat{C}_{(j)}\| = 0) \to 1$ *for any* $j \in \mathcal{A}^c$.

Theorem 2 establishes the asymptotic selection consistency in the sense that the regularized k-means clustering can eliminate the uninformative variables in the estimated cluster centers with probability tending to one. As a summary, Theorems 1 and 2 demonstrate that the proposed regularized k-means clustering is capable of performing cluster analysis and variable selection at the same time.

Note that the asymptotic estimation and selection consistency is established assuming the number of clusters K is pre-specified. When the true number of clusters is available, the asymptotic results assure that the true cluster centers and the informative variables can be accurately recovered. When the true number of clusters is not known, [23] shows the selection consistency of the number of clusters in the un-penalized clustering framework. However, it remains unclear whether similar consistent results can be obtained for the regularized methods due to the difficulty of tuning K and λ simultaneously. A numerical experiment has been conducted in section 7.2 to demonstrate the superior performance of tuning K and λ via the selection criterion in section 4.

6. Regularized model-based clustering

The regularized clustering framework can be extended to the regularized model-based clustering with the adaptive group lasso penalty. As opposed to the L_1 penalty in [16], adaptive group lasso penalty encourages the selection of variables in a factor fashion with each variable as one factor.

In general, assume each observation $X_i, i = 1, \ldots, n$ is drawn from a mixture model with $f(x) = \sum_{k=1}^{K} \pi_k f_k(x; \theta_k)$, where π_k is the mixture weight and $f_k(x; \theta_k)$ can be any distribution function of the mixture component indexed by parameter θ_k. For illustration, $f_k(x; \theta_k)$ is assumed to be a multivariate normal distribution,

$$f_k(x, \theta_k) = (2\pi)^{-p/2}|V_k|^{-1/2} \exp\left\{-\frac{1}{2}(x - C_k)^T V_k^{-1}(x - C_k)\right\}, \tag{7}$$

where $\theta_k = (C_k, V_k)$ and $|V_k|$ is the determinant of covariance matrix V_k. The regularized log-likelihood function for the observed data can be then formulated as

$$\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f_k(X_i, \theta_k)\right) - n\lambda \sum_{j=1}^{p} \frac{\|C_{(j)}\|}{\|\widetilde{C}_{(j)}\|}. \tag{8}$$

To facilitate the high-dimensional clustering as in [16], we further assume that a common diagonal covariance matrix is shared among the mixture components.

In specific, $V_k = V = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ for all k's. An EM algorithm can be employed to maximize (8), where the cluster assignment L_{ik} is treated as missing data.

If L_{ik} is available, the regularized log-likelihood function for the compete data is

$$\sum_{i=1}^{n}\sum_{k=1}^{K} L_{ik}[\log \pi_k + \log f_k(X_i;\theta_k)] - n\lambda \sum_{j=1}^{p} \frac{\|C_{(j)}\|}{\|\widetilde{C}_{(j)}\|}. \tag{9}$$

In the expectation step, the conditional expectation of (9) is denoted as

$$Q(\theta, \theta^{(t)}) = \sum_{k=1}^{K}\sum_{i=1}^{n} L_{ik}^{(t)}[\log \pi_k + \log f_k(X_i;\theta_k)] - n\lambda \sum_{j=1}^{p} \frac{\|C_{(j)}\|}{\|\widetilde{C}_{(j)}\|}. \tag{10}$$

where $L_{ik}^{(t)} = \frac{\pi_k^{(t)} f_k(X_i;\theta_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} f_k(X_i;\theta_k^{(t)})}$. In the Maximization step, maximizing (10) yields the update of the parameters,

$$\begin{aligned} \widehat{\pi}_k^{(t+1)} &= \sum_{i=1}^{n} \frac{L_{ik}^{(t)}}{n}, \\ (\widehat{\sigma}_j^2)^{(t+1)} &= \sum_{k=1}^{K}\sum_{i=1}^{n} \frac{L_{ik}^{(t)}(X_{ij} - C_{kj}^{(t)})^2}{n}. \end{aligned}$$

The centers can be obtained by a direct calculation based on the Karush-Kuhn-Tucker conditions. Specifically, for any $C_{(j)} \neq 0$,

$$\frac{\partial Q}{\partial C_{(j)}} = \frac{L^{T(t)} X_{ij} \mathbf{1}_n}{\sigma_j^2} - \frac{\text{diag}(L^{T(t)}\mathbf{1}_n)}{\sigma_j^2} - \frac{n\lambda C_{(j)}}{\|\widetilde{C}_{(j)}\|\|C_{(j)}\|}.$$

For any $C_{(j)} = 0$,

$$\|\frac{L^{T(t)} X_{ij} \mathbf{1}_n}{\sigma_j^2} - \frac{\text{diag}(L^{T(t)}\mathbf{1}_n)}{\sigma_j^2}\| \leq \frac{n\lambda}{\|\widetilde{C}_{(j)}\|}.$$

These two conditions imply that

$$\widehat{C}_{(j)}^{(t+1)} = \Big(\mathbf{I}_K - \frac{n\lambda(\sigma_j^2)^{(t+1)}(\text{diag}(L^{T(t)}\mathbf{1}_n))^{-1}}{\|\widetilde{C}_{(j)}\|\|(\text{diag}(L^{T(t)}\mathbf{1}_n))^{-1}L^{T(t)}X_{ij}\mathbf{1}_n\|}\Big)_+ \times \Big((\text{diag}(L^{T(t)}\mathbf{1}_n))^{-1}L^{T(t)}X_{ij}\mathbf{1}_n\Big),$$

where $\mathbf{I}_K$ is $K \times K$ identity matrix, $\mathbf{1}_n$ is the vector of all 1's. Note that $(A)_+$ above is component-wise, so $(A)_+ = (a_{ij+})$, where $a_{ij+} = \max(0, a_{ij})$. Therefore the element $\widehat{C}_{kj}^{(t+1)} = 0$ if

$$\lambda > \frac{\|\widetilde{C}_{(j)}\|\|(\text{diag}(L^{T(t)}\mathbf{1}_n))^{-1}L^{T(t)}X_{ij}\mathbf{1}_n\| \sum_{i=1}^{n} L_{ik}^{(t)}}{n(\sigma_j^2)^{(t+1)}}.$$

The details of the EM algorithm are as follows.

Algorithm 2 (Regularized model-based clustering).

Step 1. Initialize centers $C_1^{(0)}, \ldots, C_K^{(0)}$ by the standard k-means clustering and $\pi_k^{(0)} = \frac{1}{K}$.

Step 2. Until the termination condition is met, repeat

(a) E-step. Find $L_{ik}^{(t)} = \frac{\pi_k^{(t)} f_k(X_i;\theta_k^{(t)})}{\sum_{k=1}^K \pi_k^{(t)} f_k(X_i;\theta_k^{(t)})}$.

(b) M-step. Given $L^{(t)}$, update $\pi_k^{(t+1)}$, $(\sigma_j^2)^{(t+1)}$ and $C_{(j)}^{(t+1)}$.

Similar as Algorithm 1, the standard k-means clustering in Step 1 is randomly started multiple times to overcome its sensitivity to the initialization. The iteration in Step 2 stops when $L^{(t)}$ does not change any more.

7. Simulation study

This section examines the effectiveness of the proposed regularized k-means clustering and regularized model-based clustering, and compares them against the standard k-means and the sparse k-means. As shown by [25], the sparse k-means outperforms many other popular high-dimensional clustering algorithms in a variety of numerical experiments. To assess the performance of various clustering algorithms, the clustering error is defined as the estimated distance between an estimated clustering assignment $\hat{\psi}$ and the true assignment ψ of the sample data $X_1, \ldots, X_n$.

$$D(\hat{\psi}, \psi) = \binom{n}{2}^{-1} \left| \{(i,j) : I(\widehat{\psi}(X_i) = \widehat{\psi}(X_j)) \neq I(\psi(X_i) = \psi(X_j)); i < j\} \right|.$$

The simulated data consist of 80 observations $X_i \in R^p; i = 1, \ldots, 80$ generated as follows. First, Y_i's are uniformly sampled from $\{1,2,3,4\}$, which indicate the cluster memberships. Then for each i, the first 50 informative variables are generated from $N(\mu(Y_i), I_{50})$, where

$$\mu(Y_i) = (-\mu \mathbf{1}_{25}^T, \mu \mathbf{1}_{25}^T)^T I(Y_i = 1) + \mu \mathbf{1}_{50} I(Y_i = 2) + (\mu \mathbf{1}_{25}^T, -\mu \mathbf{1}_{25}^T)^T I(Y_i = 3) - \mu \mathbf{1}_{50} I(Y_i = 4),$$

and $\mathbf{1}_{25}$ is a vector of 25 ones, and the last $p - 50$ noise variables are generated from $N(0,1)$. To examine the clustering performance in various scenarios, we set $p = 50, 200, 500$ or 1000 and $\mu = 0.4, 0.6$ or 0.8. Clearly, the four clusters are well separated when μ is large, and can be heavily overlapped when μ is small. Furthermore, when the data dimension p increases the first 50 informative variables become harder to identify as more noise variables are present.

Two scenarios are considered. In scenario I, we focus on the clustering performance of various clustering algorithms pretending the true number of clusters is given. In scenario II, with K unknown, we compare the clustering performance of various clustering algorithms after adjusted to the tuning parameter selection. In both scenarios, the selection criterion in section 4 is used to select tuning parameters for the standard k-means, the regularized k-means clustering and the regularized model-based clustering, and gap statistic [21] is used to select tuning parameters for the sparse k-means as suggested in [25].

TABLE 1
The averaged clustering errors and their estimated standard deviations for various clustering algorithms in section 7.1

μ	Methods	p=50	p=200	p=500	p=1000
	K-means	**.085(.009)**	.193(.007)	.284(.008)	.330(.005)
0.4	Sparse k-means	.230(.021)	.212(.018)	.266(.012)	.302(.005)
	Reg. k-means	.087(.008)	**.181(.009)**	**.249(.008)**	**.296(.006)**
	Reg. model-based	.094(.005)	.196(.007)	.291(.010)	–
	K-means	.007(.002)	.025(.003)	.060(.006)	.142(.007)
0.6	Sparse k-means	.018(.003)	.016(.002)	.025(.008)	.058(.017)
	Reg. k-means	.007(.002)	**.013(.003)**	**.020(.003)**	**.044(.009)**
	Reg. model-based	**.004(.001)**	.015(.002)	.038(.006)	–
	K-means	**0(0)**	.001(.001)	.004(.002)	.015(.002)
0.8	Sparse k-means	.001(.001)	**0(0)**	.002(.001)	.004(.002)
	Reg. k-means	**0(0)**	.001(.001)	**.001(.001)**	**.001(.001)**
	Reg. model-based	**0(0)**	**0(0)**	**.001(.001)**	–

TABLE 2
The averaged numbers of selected variables and their estimated standard deviations for various clustering algorithms in section 7.1

μ	Methods	p=50	p=200	p=500	p=1000
	K-means	**50(0)**	200(0)	500(0)	1000(0)
0.4	Sparse k-means	33.3(3.05)	84.6(15.80)	127.0(39.30)	362.6(87.80)
	Reg. k-means	36.6(1.68)	**35.9(4.20)**	**45.1(9.20)**	**60.3(11.90)**
	Reg. model-based	45.2(0.61)	109.5(5.92)	98.0(15.13)	–
	K-means	**50(0)**	200(0)	500(0)	1000(0)
0.6	Sparse k-means	45.2(0.99)	128.3(9.57)	182.8(41.46)	**43.6(6.04)**
	Reg. k-means	49.8(0.12)	52.1(1.77)	**47.3(2.30)**	64.8(9.80)
	Reg. model-based	**50(0)**	**50.6(2.09)**	45.8(2.9)	–
	K-means	**50(0)**	200(0)	500(0)	1000(0)
0.8	Sparse k-means	46.4(1.09)	157.1(7.53)	126.8(30.40)	**44.9(4.41)**
	Reg. k-means	**50(0)**	**65.5(1.08)**	**53.2(1.85)**	65.3(7.03)
	Reg. model-based	**50(0)**	148.4(0.61)	56.7(3.06)	–

7.1. Scenario I: K is known

In scenario I, the number of clusters is fixed as 4 in all clustering algorithms. For all the sparse k-means, the regularized k-means clustering and the regularized model-based clustering, the tuning parameters are selected through a grid search over 20 grid points $\{10^{-2+4l/19};\ l = 0, \ldots, 19\}$. For fair comparison, the number of bootstrap samples is set as $B = 10$ in both the stability-based selection criterion in section 4 and the gap statistics, and all clustering algorithms are randomly started 100 times to overcome their dependence on the initialization. Following the setup by [25], each simulation is replicated 20 times, and the averaged clustering error and averaged number of selected informative variables are summarized in Tables 1 and 2.

Evidently, our proposed regularized k-means clustering and regularized model-based clustering deliver superior results against their competitors in terms of both clustering error and variable selection. In Table 1, the regularized k-means clustering yields smaller clustering error than both the standard k-means and the sparse k-means when $p > 50$, except that both the proposed regularized model-

based clustering and sparse k-means lead to perfect clustering when $\mu = 0.8$ and $p = 200$. When $p = 50$ with no noise variable present the regularized model-based clustering yields the best performance for $\mu = 0.6$ and 0.8, while the standard k-means clustering has great advantage for $\mu = 0.4$, whereas the performance of the sparse k-means appears to be less competitive. In Table 2, the number of selected variables by the regularized k-means clustering is much closer to the truth than that of the sparse k-means in most cases, whereas the standard k-means clustering does not performance any variable selection at all. When $p = 1000$, in the examples of $\mu = 0.6$ and $\mu = 0.8$, the regularized k-means clustering tends to include a few more variables than the sparse k-means, yet it is still reasonably close to the number of true informative variables. Furthermore, the regularized model-based clustering performs similarly as the regularized k-means clustering, but it requires substantially higher computational cost. As a consequence, the results of the regularized model-based clustering for $p = 1000$ is omitted in Tables 1 and 2 because of the long computational time.

7.2. Scenario II: K is unknown

Now we conduct a comparison of all clustering algorithms in a more realistic scenario, where the number of clusters is unknown. For illustration, we only consider $p = 200$ and $\mu = 0.8$. To select the number of clusters and tuning parameters, similar tuning procedures as in section 7.1 are applied. The grid search is conducted over $K \in \{2, \ldots, 10\}$ and the same grid points for λ as in section 7.1. The simulation is replicated 20 times and the averaged clustering errors and averaged number of selected variables are summarized in Table 3.

Again the regularization k-means clustering and regularized model-based clustering deliver superior performance in both clustering and variable selection, and outperforms the sparse k-means and the standard k-means. The performance of sparse k-means is severely deteriorated as gap statistic selects the wrong number of clusters 18 out of 20 times. The difficulty of gap statistic in selecting number of clusters is also pointed out in [25]. On the contrary, the selection criterion based on clustering stability appears to perform well in selecting the number of clusters and the tuning parameters.

To illustrate the effectiveness of the clustering stability based selection criterion, we randomly select one replication and display the estimated clustering instability and the clustering error for various values of K and λ. In Figure 1,

TABLE 3
The selected numbers of clusters, averaged numbers of selected variables, averaged clustering errors and their estimated standard deviations in section 7.2

Methods	K=2	K=4	No. selected variables	Clustering error
K-means	0	**20**	200(0)	.001(.001)
Sparse k-means	18	2	138.0(4.11)	.228(.017)
Reg. k-means	0	**20**	**50.0(0.05)**	**0(0)**
Reg. model-based	0	**20**	45.9(0.26)	.0006(.0003)

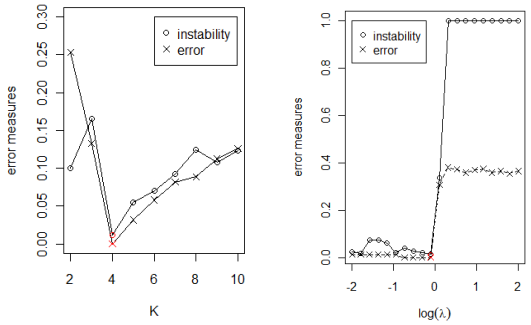

FIG 1. *The plots of clustering instability and clustering error as functions of number of clusters K and tuning parameter λ respectively.*

TABLE 4
The estimated numbers of clusters, numbers of selected variables, and clustering errors with various sample sizes in section 7.2

Sample size n	No. clusters	No. selected variables	Clustering error
20	3	86	.153
40	4	63	0
80	4	49	0

it is clear that there is a positive relevance between clustering instability and clustering error for various K or λ's.

Furthermore, we examine the behavior of the regularized k-means clustering and the tuning parameter selection criterion as sample size grows. The simulation is conducted for regularized k-means clustering with sample size $n = 20, 40, 80$. The estimated number of clusters, number of selected variables and clustering errors over 20 replications are summarized in Table 4. As sample size increases, the true number of clusters is selected with higher probability, the noninformative variables are tending not to be selected, and the clustering errors decrease implying better estimate of the clustering centers.

8. Applications to gene microarray analysis

In this section, we apply the proposed regularized k-means clustering to two benchmark microarray datasets, Leukemia [11] and Lymphoma [1]. In the Leukemia data, [11] studied microarray gene expression data to discovery two types of human acute leukemias: acute myeloid leukemia(AML) and acute lymphoblastic leukemia(ALL). This dataset consists of 72 patients in total, 25 patients with AML and 47 patients with ALL. The Gene expression levels were measured by Affymetrix microarrays containing 6817 human genes. Distinguishing ALL from AML is clinically significant for successful treatment because those chemotherapy regimens for ALL patients are different from AML patients, in which case using ALL therapy for AML (and vice versa) cases may result in

TABLE 5
The selected numbers of clusters and informative genes and clustering errors in two gene microarray examples

Data	Methods	No. clusters	No. genes	Clustering error
Leukemia	K-means	**2**	3571	**2/72**
	Sparse k-means	4	2577	**2/72**
	Reg. k-means	**2**	**211**	**2/72**
Lymphoma	K-means	2	4026	4/62
	Sparse k-means	**3**	3025	2/62
	Reg. k-means	**3**	**66**	**1/62**

distinctly reduced cute rates and possible toxicities. In the lymphoma data set, the total sample size is 62 and the number of genes is 4026. Three types of most prevalent adult lymphoid malignancies were studied: 42 cases of diffuse large B-cell lymphoma (DLBCL), 9 samples of follicular lymphoma (FL), and 11 observations of B-cell chronic lymphocytic leukemia (CLL). A specialized cDNA microarray was used to measure the gene expression levels. Both data sets are provided by [6] and available at *http* : *//stat.ethz.ch/~dettling/bagboost.html*.

Following the pre-processing steps in [8], both data sets are pre-processed by first setting a thresholding window $[100, 16000]$ and then excluding genes with $max/min \leq 5$ or $(max - min) \leq 500$. Finally a logarithmic transformation and standardization are applied. For the original lymphoma data set, some arrays contain genes with missing values. As suggested in [8], a simple 5 nearest neighbor algorithm is employed to impute the missing values.

All the clustering algorithms are randomly started 100 times to overcome their dependence on the initialization. To optimally tune the algorithms, a grid search over K and tuning parameter λ as in section 7.2 is conducted to optimize the clustering instability or gap statistic. Note that there is no true clustering assignment in both gene microarray data sets, we compare the estimated clustering assignments to the available cancer types of each tumor. The comparison results are summarized in Table 5.

In the Leukemia data, the regularized k-means clustering correctly selects 2 clusters and makes only 2 misclassification out of 72 samples. In the Lymphoma data, the regularized k-means clustering correctly selects 3 clusters and yields the smallest clustering error with only 1/62. Clearly, the regularized k-means clustering achieves competitive clustering performance with much less selected important genes compared with the sparse k-means and the standard k-means clustering algorithms. Furthermore, in the leukemia data the number of the selected important genes by the regularized k-means clustering agrees with the observations in [11, 5].

To scrutinize the performance the regularized k-means clustering in the gene microarray examples, we plot the heatmap of the Lymphoma data based on the 66 selected genes in Figure 2. The three clusters are distinct on the heatmap in that genes $1, \ldots, 22$ have significant signals in detecting FL and CLL, genes $23, \ldots, 27$ have significant signals in discriminating FL, and genes $28, \ldots, 66$ have significant signals in distinguishing DLBCL.

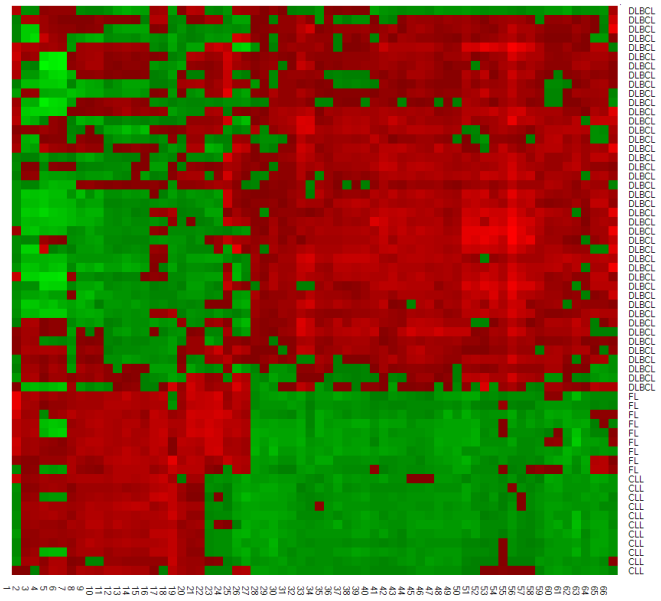

FIG 2. *Heatmap of the Lymphoma data set based on 66 genes selected by the regularized k-means clustering. Each row represents one of the 62 sample tumors and each column represents one of the 66 selected genes.*

9. Discussion

This article proposes the regularized k-means clustering which is able to simultaneously cluster high-dimensional observations and select informative variables. To optimally balance the tradeoff between model fitting and model sparsity, a tuning parameter selection criterion based on clustering stability is developed. The proposed methods deliver superior performance in both cluster analysis and variable selection, and outperform their competitors in simulated and real experiments. A possible future direction is to extend the framework of the regularized k-means clustering to other clustering algorithms, like fuzzy c-means, which relaxes the constraints of the discrete and nonnegative clustering assignment of k-means.

Appendix

Assumptions:
(i) $X_{(j)} = \bar{L}\bar{C}_{(j)} + \epsilon_{(j)}$ for $j = 1, \ldots, p$, where $\epsilon_{(j)} = (\epsilon_{1j}, \ldots, \epsilon_{nj})^T$ with ϵ_{ij} independent, $E\epsilon_{ij} = 0$ and $E\epsilon_{ij}^2 < \infty$.
(ii) The true cluster centers $\bar{C}$ is unique up to relabeling of its coordinates;
(iii) $\int \|X\|^2 P(dx) < \infty$;
(iv) Probability measure P has a continuous density f on R^p;
(v) There exists a $g(\cdot)$ such that $f(x) \leq g(\|x\|)$ and $r^p g(r)$ integrable with $r \in [0, \infty)$;

(vi) Matrix Γ defined in [18] is positive definite at $\bar{C}$;
(vii) $\arg\min_{1\le k\le K}\|X-C_k\|^2$ is unique with probability one.

Here Assumption (i) is a standard assumption for Euclidean distance based cluster analysis. Assumptions (ii)–(vi) are analogous to the assumptions in [17, 18], where p is allowed to diverge as $n\to\infty$. Assumption (vii) is necessary to prevent the ambiguity in estimating the cluster assignment matrix.

Proof of Theorem 1. First we show the estimated cluster centers $\widehat{C}_1,\dots,\widehat{C}_K$ lie in a compact region of R^p when n is large enough. It suffices to show there exists a sufficiently large closed ball $B(M)$ centered at the origin and of radius M, which contains all the estimated cluster centers when n is sufficiently large.

Note that minimization of (6) is equivalent to the minimization of

$$\int \min_{C_k\in C}\|x-C_k\|^2 P_n(dx),\quad s.t.\quad \sum_{j=1}^{p}\frac{\|C_{(j)}\|}{\|\tilde{C}_{(j)}\|}\le s_n, \tag{11}$$

where $s_n\to\infty$ as $n\to\infty$. As proved in [17], under the assumptions (ii) and (iii), there is an M_1 so large that, when n is large enough, the estimated cluster centers $\{\tilde{C}_1,\dots,\tilde{C}_K\}$ based on the standard k-means are contained in $B(M_1)$. By the fact that $s_n\to\infty$, there exists a sufficiently large s_N such that the set $\{C:\sum_{j=1}^{p}\frac{\|C_{(j)}\|}{\|\tilde{C}_{(j)}\|}\le s_N\}\supset B(M_1)$. Therefore, $B(M)=B(M_1)$ contains $\{\widehat{C}_1,\dots,\widehat{C}_K\}$ when n is sufficiently large.

Next we show that, almost surely, $W(C,P_n)-W(C,P)$ converges to zero uniformly over the subsets of $B(M)$, and then minimizing $W(\cdot,P_n)$ is asymptotically equivalent to minimizing $W(\cdot,P)$. Without loss of generality, we assume that $B(M)$ is large enough such that $\bar{C}\in B(M)$. Note that

$$\begin{aligned}
&\sup_{C\in B(M)}|W(C,P_n)-W(C,P)|\\
&=\sup_{C\in B(M)}\Big|\int\min_{C_k\in C}\|x-C_k\|^2P_n(dx)+\sum_{j=1}^{p}\lambda_j\|C_{(j)}\|-\int\min_{C_k\in C}\|x-C_k\|^2P(dx)\Big|\\
&\le\sup_{C\in\ B(M)}\Big|\int\min_{C_k\in C}\|x-C_k\|^2P_n(dx)-\int\min_{C_k\in C}\|x-C_k\|^2P(dx)\Big|\\
&\qquad\qquad+\sup_{C\in B(M)}\sum_{j=1}^{p}\lambda_j\|C_{(j)}\|.
\end{aligned}$$

The first term converges almost surely to zero because $p=o(n)$ and the uniform SLLN of standard k-means [17], and the second term converges almost surely to zero because $\|C_{(j)}\|$ and $\|\hat{C}_{(j)}\|$ are bounded on $B(M)$ and $n^{1/2}\lambda p\to 0$.

Finally we prove that $\|\widehat{C}-\bar{C}\|=O_p(n^{1/2}\lambda p^{-1})$. Denote an empirical process $G_n(\cdot)=n^{1/2}(P_n(\cdot)-P(\cdot))$ and denote $\phi(x,C)=\min_{1\le k\le K}\|x-C_k\|^2$. Then

$$W(\widehat{C},P_n)=P_n\phi(\cdot,\widehat{C})+\sum_{j=1}^{p}\lambda_j\|\widehat{C}_{(j)}\|=P\phi(\cdot,\widehat{C})+n^{-1/2}G_n\phi(\cdot,\widehat{C})+\sum_{j=1}^{p}\lambda_j\|\widehat{C}_{(j)}\|.$$

Therefore under the conditions (iii) and (iv), Lemma D in [18] implies that

$$\begin{aligned} W(\widehat{C}, P_n) &= W(\bar{C}, P_n) - n^{-1/2} Z_n^T (v(\widehat{C}) - v(\bar{C})) \\ &\quad + \frac{1}{2}(v(\widehat{C}) - v(\bar{C}))^T \Gamma (v(\widehat{C}) - v(\bar{C})) + \sum_{j=1}^{p} \lambda_j \|\widehat{C}_{(j)}\| \\ &\quad - \sum_{j=1}^{p} \lambda_j \|\bar{C}_{(j)}\| + o_p(n^{-1/2} r_n) + o_p(r_n^2). \end{aligned}$$

where $r_n = \|\widehat{C} - \bar{C}\|$, $v(\widehat{C})$ and $v(\bar{C})$ are the vectorized $\widehat{C}$ and $\bar{C}$, $Z_n \in R^{kp}$ is asymptotically $N(0, V)$ with $V \in R^{kp \times kp}$ as defined in Lemma D of [18]. By the definition of $\widehat{C}$, $W(\widehat{C}, P_n) \leq W(\bar{C}, P_n)$. Therefore,

$$\begin{aligned} &-n^{-1/2} Z_n^T (v(\widehat{C}) - v(\bar{C})) + \frac{1}{2}(v(\widehat{C}) - v(\bar{C}))^T \Gamma (v(\widehat{C}) - v(\bar{C})) \\ &\quad + \sum_{j=1}^{p} \lambda_j (\|\widehat{C}_{(j)}\| - \|\bar{C}_{(j)}\|) + o_p(n^{-1/2} r_n) + o_p(r_n^2) \leq 0. \end{aligned}$$

Assumption (vi) guarantees that $(v(\widehat{C}) - v(\bar{C}))^T \Gamma (v(\widehat{C}) - v(\bar{C})) = O_p(p r_n^2)$, and the fact that the elements of Z_n are in the order of $O_p(1)$ implies that $-n^{-1/2} Z_n^T (v(\widehat{C}) - v(\bar{C})) = O_p(n^{-1/2} p^{1/2} r_n)$. By the fact that $\lambda_j = \lambda \|\widetilde{C}_{(j)}\|^{-1}$ and central limit theorem of $\tilde{C}$, which is extended from standard k-means clustering in [18], $\sum_{j=1}^{p} \lambda_j (\|\widehat{C}_{(j)}\| - \|\bar{C}_{(j)}\|) = O_p(n^{1/2} \lambda r_n)$. According to the assumption $n^{-2} \lambda^{-2} p \to 0$,

$$O_p(p r_n^2) + O_p(n^{1/2} \lambda r_n) \leq O_p(n^{-1/2} p^{1/2} r_n) + o_p(n^{-1/2} r_n) + o_p(r_n^2)$$

implies that $r_n = O_p(n^{1/2} \lambda p^{-1})$. □

Proof of Theorem 2. We only prove $P(\|\widehat{C}_{(p)}\| = 0) \to 1$ by contradiction, and similar treatment can yield that $P(\|\widehat{C}_{(j)}\| = 0) \to 1$, for all $p_0 + 1 \leq j \leq p - 1$. If $\widehat{C}_{(p)} \neq 0$, then $\|\widehat{C}_{(p)}\|$ is differentiable with respect to its components. Karush-Kuhn-Tucker (K.K.T.) condition implies that

$$\begin{aligned} 0 &= -\frac{2}{\sqrt{n}} \widehat{L}^T (X_{(p)} - \widehat{L}\widehat{C}_{(p)}) + \sqrt{n} \lambda_p \frac{\widehat{C}_{(p)}}{\|\widehat{C}_{(p)}\|} \\ &= -\frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)(X_{(p)} - \widehat{L}\widehat{C}_{(p)}) - \frac{2}{\sqrt{n}} \bar{L}^T (X_{(p)} - \widehat{L}\widehat{C}_{(p)}) + \sqrt{n} \lambda_p \frac{\widehat{C}_{(p)}}{\|\widehat{C}_{(p)}\|} \\ &= -\frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)(X_{(p)} - \bar{L}\widehat{C}_{(p)}) + \frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)(\widehat{L} - \bar{L})\widehat{C}_{(p)} \\ &\quad - \frac{2}{\sqrt{n}} \bar{L}^T (X_{(p)} - \bar{L}\widehat{C}_{(p)}) + \frac{2}{\sqrt{n}} \bar{L}^T (\widehat{L} - \bar{L})\widehat{C}_{(p)} + \sqrt{n} \lambda_p \frac{\widehat{C}_{(p)}}{\|\widehat{C}_{(p)}\|} \\ &= \frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)\bar{L}\widehat{C}_{(p)} - \frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)\epsilon_{(p)} + \frac{2}{\sqrt{n}} (\widehat{L}^T - \bar{L}^T)(\widehat{L} - \bar{L})\widehat{C}_{(p)} \\ &\quad + \frac{2}{\sqrt{n}} \bar{L}^T (\widehat{L} - \bar{L})\widehat{C}_{(p)} - \frac{2}{\sqrt{n}} \bar{L}^T \epsilon_{(p)} + \Big(\frac{2}{n} \bar{L}^T \bar{L} + \frac{\lambda_p}{\|\widehat{C}_{(p)}\|} I \Big) \sqrt{n} \widehat{C}_{(p)}. \end{aligned}$$

Note that $\widehat{C} \to \bar{C}$ in Theorem 1 together with Assumption (vii) implies that the estimated cluster assignment matrix $\widehat{L}$ converges in probability to the true cluster assignment matrix $\bar{L}$. Therefore, in the last equality, the first four terms are of the order $o_p(1)$, and the fifth term is of the order $O_p(1)$ due to assumption (i). It follows from the fact that $\frac{2}{n}\bar{L}^T\bar{L}$ is a nonnegative matrix and the component-wise central limit theorem of standard k-means that

$$\left\| \left(\frac{2}{n}\bar{L}^T\bar{L} + \frac{\lambda_p}{\|\widehat{C}_{(p)}\|} I \right) \sqrt{n}\widehat{C}_{(p)} \right\| \geq \left\| \frac{\lambda_p}{\|\widehat{C}_{(p)}\|} \sqrt{n}\widehat{C}_{(p)} \right\| = \frac{\sqrt{n}\lambda}{\|\widetilde{C}_{(p)}\|} = O(n\lambda).$$

Note that $n\lambda \to \infty$ according to the assumption $n^{-2}\lambda^{-2}p \to 0$. So the last term diverges to infinity and dominates the first five terms, which leads to the contradiction to the above K.K.T. condition. Therefore, $\widehat{C}_{(p)}$ must be equal to 0 with probability tending to one. This completes the proof. □

References

[1] Alizadeh, A., Eisen, M., Davis, R., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., Powell, J., Yang, L., Marti, G., Moore, T., Hudson, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, R., Chan, W., Greiner, T., Weisenburger, D., Armitage, Warnke, R., Levy, R., Wilson, W., Grever, M., Byrd, J., Botstein, D., Brown, P., and Staudt, L. (2000), "Different Types of Diffuse Large B-cell Lymphoma Identified by Gene Expression Profiling," *Nature*, **403**, 503-511.

[2] Bansal, N., Blum, A., and Chawla, S. (2004), "Correlation Clustering," *Machine Learning*, **56**, 86-113.

[3] Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002), "A Stability Based Method for Discovering Structure in Clustered Data," *Pacific Symposium on Biocomputing*, 6-17.

[4] Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, **37**, 373-384. MR1365720

[5] Changdra, B., Shanker, S., and Mishra, S. (2006), "A New Approach: Interrelated Two-way Clustering of Gene Expression Data," *Statistical Methodology*, **3**, 93-102. MR2210635

[6] Dettling, M. (2004), "BagBoosting for Tumor Classification with Gene Expression Data," *Bioinformatics*, **20**, 3583-3593.

[7] Donoho, D., and Jin, J. (2008), "Higher Criticism Thresholding: Optimal Feature Selection When Useful Features are Rare and Weak," *The Proceedings of the National Academy of Sciences of the United States of America*, **105**, 14790-14795.

[8] Dudoit, S., Fridlyand, J., and Speed, T. (2002), "Comparision of Discrimination Methods for the Classification of Tumor Using Gene Expression Data," *Journal of the American Statistical Association*, **97**, 77-87. MR1963389

[9] Fan, J. and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, **32**, 928-961. MR2065194

[10] Fang, Y. and Wang, J. (2011), "Penalized Cluster Analysis with Applications to Family Data," *Computational Statistics and Data Analysis*, **55**, 2128-2136. MR2785119

[11] Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, A., Bloomfield, C., and Lander, E. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, **286**, 531-537.

[12] Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Pairwise Variable Selection for High-dimensional Model-based Clustering," *Biometrics*, **66**, 793-804. MR2758215

[13] Hall, P., Marron, J.S., and Neeman, A. (2005), "Geometric Representation of High Dimension, Low Sample Size Data," *Journal of the Royal Statistical Society, Series B*, **67**, 427-444. MR2155347

[14] Lloyd, S.P. (1982), "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, **28**, 129-137. MR0651807

[15] MacQueen, J. (1967), "Some Methods for Clasification and Analysis of Multivariate Observations," *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297. MR0214227

[16] Pan, W., and Shen, X. (2007), "Penalized Model-Based Clustering with Application to Variable Selection," *Journal of Machine Learning Research*, **8**, 1145-1164.

[17] Pollard, D. (1981), "Strong Consistency of K-means Clustering," *The Annals of Statistics*, **9**, 135-140. MR0600539

[18] Pollard, D. (1982), "A Central Limit Theorem for K-means Clustering," *The Annals of Probability*, **10**, 919-926. MR0672292

[19] Raftery, A., and Dean, N. (2006), "Variable Selection for Model-based Clustering," *Journal of the American Statistical Association*, **101**, 168-178. MR2268036

[20] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, **58**, 267-288. MR1379242

[21] Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Data Set via the Gap Statistic," *Journal of the Royal Statistical Society, Series B*, **63**, 411-423. MR1841503

[22] Wang, H., and Leng, C. (2008), "A Note on Adaptive Group Lasso," *Computational Statistics and Data Analysis*, **52**, 5277-5286. MR2526593

[23] Wang, J. (2010), "Consistent Selection of the Number of Clusters via Cross Validation," *Biometrika*, **97**, 893-904. MR2746159

[24] Wang, S., and Zhu, J. (2008), "Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data," *Biometrics*, **64**, 440-448. MR2432414

[25] Witten, D., and Tibshirani, R. (2010), "A Framework for Feature Selection in Clustering," *Journal of the American Statistical Association*, **105**, 713-726. MR2724855

[26] Xie, B., Pan, W., and Shen, X. (2008), "Variable Selection in Penalized Model-Based Clustering Via Regularization on Grouped Parameters," *Biometrics*, **64**, 921-930. MR2526644

[27] Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B*, **68**, 49-67. MR2212574

[28] Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, **101**, 1418-1429. MR2279469

[29] Zou, H., and Zhang, H. (2009), "On the Adaptive Elastic-net with A Diverging Number of Parameters," *The Annals of Statistics*, **37**, 1733-1751. MR2533470