

# SPARK: A New Clustering Algorithm for Obtaining Sparse and Interpretable Centroids



Naoto Yamashita and Kohei Adachi

**Abstract** *k*-means clustering is one of the popular procedures for multivariate analysis in which observations are classified into a reduced number of clusters. The resulting centroid matrix is referred to capture variables which characterize clusters, but between-clusters contrasts in the centroid matrix are not always clear and thus difficult to interpret. In this research, we address the problem in interpretation and propose a new procedure of *k*-means clustering which produces a sparse and thus interpretable centroid matrix. The proposed procedure is called SPARK. In SPARK, the sparseness of the centroid matrix is constrained and therefore it contains a number of exact zero elements. Because of this, the contrasts between-clusters are highlighted and it allows us to interpret clusters easier in comparison with the standard *k*-means clustering. A sparsity selection procedure for determining the optimal sparsity of the centroid with reduced computational load is also proposed. Behaviors of the proposed procedure are evaluated by two real data examples, and the results indicate that SPARK performs well for dealing with real world problems.

**Keywords** *k*-means clustering · Sparse estimation · Interpretability

## 1 Introduction

*k*-means clustering, known as a non-hierarchical clustering procedure, is widely used for extracting the homogeneity of observations, by assigning them into a small number of clusters. Let  $\mathbf{X}$  be an  $n$ -observations  $\times$   $p$ -variables matrix, and the *k*-means clustering is formulated as a minimization of the least squares loss function defined as

$$f(\mathbf{M}, \mathbf{Y}) = \sum_{i,l} m_{il} \|\mathbf{x}'_{(i)} - \mathbf{y}_l\|^2 = \|\mathbf{X} - \mathbf{M}\mathbf{Y}'\|^2, \quad (1)$$

---

N. Yamashita (✉) · K. Adachi  
Graduate School of Human Sciences, Osaka University, 1-2 Yamadaoka, Suita, Osaka, Japan  
e-mail: nyamashita@hus.osaka-u.ac.jp; nyamashita.hus@osaka-u.ac.jp

K. Adachi  
e-mail: adachi@hus.osaka-u.ac.jp

where  $\mathbf{M} = \{m_{il}\}$  is an  $n$ -observations  $\times$   $p$ -variables membership matrix and  $\mathbf{Y} = \{y_{jl}\}$  is a  $p$ -variables  $\times$   $k$ -clusters centroid matrix.  $\mathbf{x}_{(i)}$  and  $\mathbf{y}_l$  denote the  $i$ th row vector and the  $l$ th column vector of  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively.

The centroid matrix is referred for interpreting what variables characterize the clusters, and the within- and between-clusters contrasts in the centroid matrix are of help for the interpretation. These contrasts, however, are not always clearly observed and therefore the interpretation is difficult, as exemplified in Sect. 4. A typical strategy to discriminate the clusters is to replace the elements close to zero in the centroid matrix with zeros, by a certain threshold. It is not recommended, however, in that the threshold totally depends on users' decision, and it can spoil the reliability of the interpretation and the following decisions.

In this article, considering the above problem in interpretability of the resulting centroid matrix, we propose a new algorithm for clustering which produces an easily interpreted centroid matrix. We call this algorithm SPARK (abbreviation of Sparse  $k$ -means). In SPARK, the resulting centroid matrix is sparse in that it contains a number of entries exactly equal to zero. The contrasts of the clusters are therefore emphasized, without any subjective threshold, which facilitates the easier and more coherent interpretation than the existing procedures. Such a centroid matrix is obtained by minimizing (1) subject to the constraint that  $\mathbf{Y}$  has a specific number of zero elements, namely,

$$Sp(\mathbf{Y}) = r \quad (2)$$

where  $Sp(\mathbf{Y})$  is the number of zero in  $\mathbf{Y}$ . The positive integer  $r$  is specified beforehand.

### 1.1 Related Procedure

Sun et al. (2012) proposed regularized  $k$ -means clustering for obtaining such sparse centroid matrix, which is similar to the proposed method. It is formulated as a minimization of (1) subject to the row-wise constraint on  $\mathbf{Y}$

$$\|\mathbf{y}_{(j)}\| \leq \lambda_j \quad (j = 1, \dots, p) \quad (3)$$

where  $\|\mathbf{y}_{(j)}\|$  is an  $L_1$ -norm of the  $\mathbf{Y}$ 's  $j$ th row vector  $\mathbf{y}_{(j)}$  and a tuning parameters  $\lambda_j$  ( $j = 1, \dots, p$ ) control the resulting sparsity of  $\mathbf{Y}$ . It therefore contains a number of zero elements, since the  $L_1$ -norm of rows of  $\mathbf{Y}$  is constrained to be less than  $\lambda_1, \dots, \lambda_p$ . This minimization is equivalent to the minimization of the following function;

$$f(\mathbf{M}, \mathbf{Y}) + \sum_j^p \lambda_j \|\mathbf{y}_{(j)}\|. \quad (4)$$

We call this approach as a *penalty approach*, in that it adds the penalty function  $\sum_j^p \lambda_j ||\mathbf{y}_{(j)}||$  to the original loss function (1). The tuning parameters take any positive integer, which are commonly determined by cross-validation. Similar approaches can be found in Witten and Tibshirani (2010) and Hastie et al. (2015). Penalty approach is originally proposed for avoiding over-fitting in clustering. Generalizability, however, does not always results in the easier interpretability of the clusters, which we focus on in this article. The proposed procedure directly controls the number of zero elements  $r$  in the centroid matrix within a restricted range, without introducing tuning parameters as in the penalty approaches. Within- and between-contrasts in the centroid matrix are therefore highlighted, and it allows users to find what variables manifest the clusters easily. It should be noted that controlling  $r$  cannot consider all possible values of  $\lambda_1, \dots, \lambda_p$ . For interpretation of clusters, however, inspecting all possible  $\lambda$ s is not necessary, and sparseness of  $\mathbf{Y}$  can be determined by how many elements in  $\mathbf{Y}$  are zero and ignorable.

## 2 Algorithm

The proposed procedure SPARK is formulated as the following constrained minimization problem;

$$\min_{\mathbf{M}, \mathbf{Y}} f(\mathbf{M}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 \tag{5}$$

subject to the sparsity constraint (2) and the membership constraint is imposed on  $\mathbf{M}$  such that

$$m_{il} \in \{0, 1\} \text{ and } \sum_l m_{il} = 1. \tag{6}$$

The parameter matrices are alternately and iteratively updated in the M-step and Y-step, respectively, starting from multiple sets of initial values in order to avoid accepting a local minimum as the final solution. In these steps, the current parameter matrix is replaced by the one minimizing (1) keeping the other parameter matrix fixed. The update formulae used in the M-step and Y-step are presented as follows. *M-step* The minimization of  $f(\mathbf{M}, \mathbf{Y})$  with fixed  $\mathbf{Y}$  subject to (6) is achieved by the  $k$ -means algorithm with the fixed centroid (MacQueen 1967). Therefore, the optimal  $\mathbf{M} = \{m_{il}\}$  is obtained by

$$m_{il} = \begin{cases} 1 & (l = \arg \min_l f(\mathbf{M}, \mathbf{Y})) \\ 0 & (\text{otherwise}) \end{cases}, \tag{7}$$

for  $i = 1, \dots, n$ .

*Y-step* Using the matrix  $\mathbf{C} = \mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}$ , (1) is rewritten as

$$\begin{aligned}
f(\mathbf{M}, \mathbf{Y}) &= \|\mathbf{X} - \mathbf{M}\mathbf{Y}'\|^2 \\
&= \|\mathbf{X} - \mathbf{M}\mathbf{C}' + \mathbf{M}\mathbf{C}' - \mathbf{M}\mathbf{Y}'\|^2 \\
&= \|\mathbf{X} - \mathbf{M}\mathbf{C}'\|^2 + \|\mathbf{D}^{1/2}(\mathbf{C} - \mathbf{Y})\|^2 \\
&\quad - \text{tr}(\mathbf{X} - \mathbf{M}\mathbf{C}')'(\mathbf{M}\mathbf{C}' - \mathbf{M}\mathbf{Y}'). \tag{8}
\end{aligned}$$

where  $\mathbf{D} = \text{diag}\{d_{11}, \dots, d_{ll}, \dots, d_{kk}\}$  denotes the  $k \times k$  diagonal matrix whose  $l$ th diagonal element is equal to the number of the observations classified into the  $l$ th cluster ( $l = 1, \dots, k$ ). The third term is proved to be zero as follows;

$$\begin{aligned}
&\text{tr}(\mathbf{X} - \mathbf{M}\mathbf{C}')'(\mathbf{M}\mathbf{C}' - \mathbf{M}\mathbf{Y}') \\
&= \text{tr}\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X} - \text{tr}\mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{X} - \text{tr}\mathbf{X}'\mathbf{M}'\mathbf{M} + \text{tr}\mathbf{X}'\mathbf{M}'\mathbf{M} \\
&= 0. \tag{9}
\end{aligned}$$

Therefore, minimizing the second term in (8),  $g(\mathbf{Y}) = \|\mathbf{D}^{1/2}(\mathbf{C} - \mathbf{Y})\|^2$ , is equivalent to the minimization of  $f(\mathbf{M}, \mathbf{Y})$  with respect to  $\mathbf{Y}$ . Further,  $g(\mathbf{Y})$  is rewritten as

$$g(\mathbf{Y}) = \sum_{(j,l) \in Z} d_{ll}^{1/2} c_{jl}^2 + \sum_{(j,l) \in Z^\perp} d_{ll}^{1/2} (c_{jl} - y_{jl})^2 \geq \sum_{(j,l) \in Z} d_{ll}^{1/2} c_{jl}^2 \tag{10}$$

where the  $Z$  denotes  $r$  pairs of indices  $(j, l)$ s indicating the locations of  $y_{jl}$ s to be zero. The last equality holds when the second term in (10) is equal to zero, that is, when  $y_{jl}$  with  $(j, l) \in Z^\perp$  is taken equal to the corresponding  $c_{jl}$ . In addition, the limit  $\sum_{(j,l) \in Z} d_{ll}^{1/2} c_{jl}^2$  is minimal when  $Z$  is composed of the indices of the  $r$  smallest  $c_{jl}^2$ s among all squared elements in  $\mathbf{C}$ . Therefore,  $\mathbf{Y}$  that minimizes  $g(\mathbf{Y})$  is obtained as

$$y_{jl} = \begin{cases} 0 & (\text{iff } c_{jl}^2 \leq c_{[r]}^2) \\ c_{jl} & (\text{otherwise}) \end{cases} \tag{11}$$

for  $l = 1, \dots, k$  and  $j = 1, \dots, p$ , where  $c_{[r]}^2$  denotes the  $r$ th smallest value among all  $c_{jl}^2$ s. The update formulae (7) and (11) are used in the M-step and Y-step, respectively, and it is guaranteed that function value of  $f(\mathbf{M}, \mathbf{Y})$  monotonically decreases in each of these steps. As presented in this section,  $\mathbf{M}$  and  $\mathbf{Y}$  are alternately updated until the convergence is reached. In the following real data examples, we used 100 different initial values for  $\mathbf{M}$  and  $\mathbf{Y}$ .

### 3 Sparsity Selection Based on Information Criteria

In the proposed procedure, the number of zeros in  $\mathbf{Y}$  has to be specified as a positive integer  $r$  in (2). In this article, the minimum and maximum of  $r$ ,  $r_{\min}$ ,  $r_{\max}$ , are defined as

$$r_{min} = 1, \quad r_{max} = p \times (k - 1) \quad (12)$$

considering that  $\mathbf{Y}$  has  $p$  non-zero elements when  $\mathbf{Y}$  has a perfect cluster structure; each variable is associated with only one cluster. Selecting the number of zero elements in  $\mathbf{Y}$  can be considered as a model selection problem, since this selection partially specifies the model part of  $\mathbf{MY}'$  fitted to  $\mathbf{X}$ . In this respect, the information criterion such as AIC and BIC is suitable for specifying  $r$ , which controls how sparse the model is to be fitted to the data. In this section, we propose two criteria in order to select the “best”  $r$  among the interval  $[r_{min}, r_{max}]$ .

Here, let  $\mathbf{E} = \{e_{ij}\}$  be the matrix of errors defined as  $\mathbf{E} = \mathbf{X} - \mathbf{MY}'$ . Under the assumption that  $\mathbf{X}$  is generated by  $\mathbf{X} = \mathbf{MY}' + \mathbf{E}$  with  $e_{ij}$  distributed independently and identically according to  $N(0, \sigma^2)$  for all  $i$ s and  $j$ s with a specific error variance  $\sigma^2$ , it can be shown that the least squares estimation and maximum likelihood estimation in SPARK are equivalent. The log-likelihood function to be maximized in the ML estimation is

$$l(\mathbf{M}, \mathbf{Y}) = -\frac{np}{2} \log \|\mathbf{X} - \mathbf{MY}'\|^2 \quad (13)$$

including  $f(\mathbf{M}, \mathbf{Y})$  to be minimized in the least square estimation. With an arbitrary  $r$ , the maximum of  $l(\mathbf{M}, \mathbf{Y})$  is attained as

$$l(\mathbf{M}, \mathbf{Y}) \leq -\frac{np}{2} \log f_{min}(r). \quad (14)$$

where  $f_{min}(r)$  denotes the attained function value of (1). By (14), the information criteria  $AIC(r)$  and  $BIC(r)$  with the specific  $r$  are obtained by

$$AIC(r) = np \times \log f_{min}(r) + 2v(r) \quad (15)$$

$$BIC(r) = np \times \log f_{min}(r) + \log(np) \times v(r) \quad (16)$$

where  $v(r)$  denotes the number of parameter to be estimated with a certain  $r$ ;

$$v(r) = n + kp - r, \quad (17)$$

Therefore,  $r$  can be determined by  $r = \arg \min_{r_{min} \leq r \leq r_{max}} AIC(r)$  or  $BIC(r)$  in terms of minimizing the model selection criteria. This approach is considered to be computationally inefficient, however, as of 100 run of SPARK are required, in order to avoid a local minimum, for each of all possible  $r$ s. When  $\mathbf{X}$  is of a large size, ( $\mathbf{X}$  contains many observations and variables) the resulting centroid matrix is also of a large size, and thus higher computational cost is required for each run.

In order to find such  $r$  with lower computational cost, we propose the following algorithm.

Step 1. Set  $S_{initial}$  and  $S_{decrease}$  to an integer within the range  $[0, 1]$ . Set  $r_t = S_{initial} \times r_{max}$

Step 2. Repeat Step 3 to Step 4 while  $S > 1$ .

Step 3. (*Forward search*) Repeat (a) to (c).

(a) Set  $r = r_t$  and compute

$$\Delta AIC(r) = AIC(r + 1) - AIC(r) \quad (18)$$

or

$$\Delta BIC(r) = BIC(r + 1) - BIC(r) \quad (19)$$

(b) If  $\Delta AIC(r)$  or  $\Delta BIC(r)$  is smaller than 0, set  $r_t = r_t + S$  and go back to 2. Otherwise proceed to (c).

(c) Set  $S = S \times S_{decrease}$  and proceed to the *backward search*.

Step 4. (*Backward search*) Repeat (a) to (c).

(a) Set  $r = r_t$  and compute  $\Delta AIC(r)$  or  $\Delta BIC(r)$ .

(b) If  $\Delta AIC(r)$  or  $\Delta BIC(r)$  is greater than 0, set  $r_t = r_t - S$  and go back to 4. Otherwise proceed to (c).

(c) Set  $S = S \times S_{decrease}$  and proceed to the *forward search*.

Step 5. If the previous step is *Forward search*, repeat *backward search* with  $S = 1$  until  $\Delta AIC(r)$  or  $\Delta BIC(r)$  is positive; otherwise repeat *Forward search* until  $\Delta AIC(r)$  or  $\Delta BIC(r)$  is negative.

The above algorithm seeks  $r$  which minimizes  $AIC(r)$  or  $BIC(r)$  within the range  $[r_{min}, r_{max}]$  by repeating the forward and backward search and reducing the step size  $S$  at each step of the iteration, starting from the initial step size  $r_{max} \times S_{initial}$ . The rate of decrement of the step size is controlled by  $S_{decrease}$ . The total computational cost is therefore dramatically reduced compared with applying SPARK for computing  $AIC(r)$  or  $BIC(r)$  for all  $rs$ . In the following simulation and the real data examples, we set  $S_{initial} = 0.9$  and  $S_{decrease} = 0.7$  which is empirically confirmed to be well-performed.

## 4 Real Data Examples

In this section, we demonstrate that SPARK extracts the sparse centroids underlying the dataset and facilitates interpretation of the centroid, with keeping the correctness of classification.

### 4.1 Example 1: Fisher’s Iris Data

In the first example, SPARK was applied to Fisher’s Iris data, where 150 samples, which are originally sampled from three species, were measured with respect to four variables. In order to find the optimal sparsity, the sparsity selection procedure based on BIC was used. It suggested that  $r = 2$  was the best, and we also applied the standard  $k$ -means clustering to Iris data for comparison.

The estimated centroids are shown in Table 1 as a heatmap. As found in Table 1, the contrast between the first (C1) and the second (C2) clusters can be seen in Sepal.Length and Sepal.Width. In addition, C2 is different from the rest of clusters with respect to Sepal.Width The contingency table of two partitions, the species of samples and the estimated membership, for SPARK and the one for  $k$ -means, are shown in Table 2. It can be seen that the estimated memberships correspond to the species, in that  $(49 + 37 + 42)/150 = 85.3\%$  of the observations are correctly classified, while  $(50 + 39 + 36)/150 = 89.2\%$  in the  $k$ -means. These results indicate that SPARK appropriately produces sparser and thus easy-to-interpret centroid matrix in comparison with the exiting method, keeping the accuracy of classification.

**Table 1** Estimated centroid matrices by SPARK for Fisher’s iris dataset with  $r = 2$  and  $k$ -means clustering

		Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
SPARK	C1	1.065		0.966	0.999
	C2		-0.928	0.322	0.236
	C3	-1.011	0.850	-1.301	-1.251
$k$ -means	C1	1.132	0.088	0.993	1.014
	C2	0.050	-0.880	0.347	0.281
	C3	-1.011	0.850	-1.301	-1.251

**Table 2** Contingency table for species versus the estimated partitions by SPARK and  $k$ -means

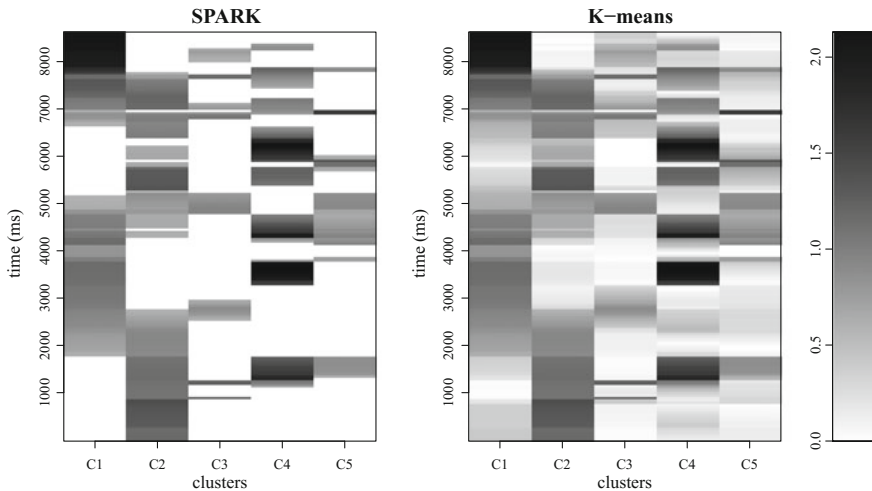
	SPARK			$k$ -means		
	C1	C2	C3	C1	C2	C3
Setosa	49	0	0	50	0	0
Versicolor	1	37	8	0	39	14
Virginica	0	13	42	0	11	36

## 4.2 Example 2: Vicon Physical Action Dataset

The second example is Vicon Physical Action Dataset (Lichman 2013). A subject's walking was recorded by the 3-axis motion sensors attached to the subject's right and left wrists, elbows, knees, and ankles. The activity was recorded for approximately 8000 ms with the frequency of 20 Hz. Therefore we have 24 ( $x$ -/ $y$ -/ $z$ -axis sensors of right and left wrists, elbows, knees and ankles)  $\times$  173 (time elapsed) data matrix.  $k$ -means clustering is applied to the data matrix and the resulting centroid matrix is shown in Fig. 1 as a heatmap. The number of clusters is set to 5 which explains 75% of the total variance of the dataset.

We can interpret the estimated five clusters by referring the  $173 \times 5$  centroid matrix as follows. For example, the first (C1) and the second (C2) clusters are well discriminated against the others; the first cluster is characterized by the lower output value in the middle phase of records (around 2000–6000 ms) and the higher value in the latter phase (around 6000–8500 ms), while this variation in the sensor outputs is shifted for 2000 ms earlier in the second cluster. The third (C3), fourth (C4) and fifth (C5) clusters are, however, hard to be discriminated mutually, in that the time evolutions of values are similar to each other especially in the early phase.

Before applying SPARK to the dataset, the sparsity selection procedures were applied. The AIC- and BIC-based procedures suggested that  $r = 332$  and  $r = 461$  were the best, respectively. We therefore determined to set  $r = 461$  in order to obtain the sparser centroid matrix. This means that approximately 53.3% of the all elements of the centroid matrix were estimated as zero. The number of clusters was set at 5, as in the example of the  $k$ -means clustering in Sect. 1.



**Fig. 1** Estimated centroid matrix by SPARK with  $r = 461$  and  $k$ -means (absolute transformed) for Vicon Physical Action Dataset



**Table 3** Estimated membership of 24 sensors; x/y/z-axis sensor on the right (R) and left (L) wrist, elbow, knee and ankle

	wrist		elbow		knee		ankle	
	L	R	L	R	L	R	L	R
C1		z		z				
C2						z		z
C3	y/z	y	y/z	y	y	y	y	y
C4	x	x	x	x	x	x	x	x
C5					z		z	

The resulting centroid matrix is represented as a heatmap in Fig. 1. The elements estimated as zero are colored in white. It can be seen that, compared with the standard *k*-means clustering, the estimated centroid is sparse enough and the contrasts between clusters are clearer than in the *k*-means clustering solution. Based on the sparse centroid, each cluster can be interpreted as follows; the sensors classified into the first cluster show the lower values from approximately 2000–5000 ms and the higher values from 6500 ms to the end of recording, and this variation of sensor outputs is earlier by 1500 ms in the second cluster. The third cluster is characterized by the lower values around 6000 ms, which makes the cluster different from the other clusters. In the fourth cluster, the lower values and the higher values alternately appear except in the early phase of recording, while the sensor outputs are almost stable in the fifth cluster.

The centroids obtained by *k*-means are less sparse than the centroids for SPARK and the characteristics of clusters are unclear. As a measure of interpretability, Lorenzo-Seva (2003) proposed the index of simplicity called LS index in the context of factor analysis. The LS index ranges from 0 (least simple) to 1 (most simple) and the values LS index for the centroid matrices were 0.313 in the *k*-means and 0.590 in the SPARK, which indicates the sparsely estimated centroids are more simple and thus more interpretable compared with the existing method.

The sensor classified into each cluster are shown in Table 3. The first cluster is composed of the z-axis sensors on the right arm, while those on the left arm are classified into the third cluster. It indicates that the subject’s horizontal movement in the left and right arms are expressed in the first and the second clusters. The third cluster is composed of 10 sensors, the y/z-axis sensors on the left arm and the y-axis sensors on the leg. The x-axis sensor on all parts are classified into the fourth cluster, and refereeing the sparse centroids in Fig. 1 therefore indicate that the clear difference between the x-axis and the y-axis movement is observed around 6000 ms.

## 5 Concluding Remarks

In this article, we proposed a new procedure of clustering called SPARK, which produces a sparse centroid matrix. The interpretation of the centroid matrix is easier compared with the ordinal  $k$ -means clustering by the sparsity constraint imposed on the centroid matrix. It is also possible to obtain such sparse centroid by adding a penalty term to the loss function of  $k$ -means clustering, as proposed by some authors. These procedures mainly aim to improve the robustness of clustering through the sparse estimation of centroid matrix. In SPARK, on the other hand, we rather focus on the interpretability of the resulting centroid matrix than robustness. The sparseness of the centroid matrix is therefore controlled by the number of zero elements in the centroid matrix, which is closely related to its interpretation. The results of the two real data examples indicate that the estimated sparse centroids surely facilitate to capture the characteristic of the clusters.

## References

- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press.
- Lichman, M. (2013). UCI machine learning repository. <http://archive.ics.uci.edu/ml>
- Lorenzo-Seva, U. (2003). A factor simplicity index. *Psychometrika*, 68(1), 49–60.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281–297.
- Sun, W., Wang, J., & Fang, Y. (2012). Regularized  $k$ -means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6, 148–167.
- Witten, D., & Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490), 713–726.