Journal of Classification (2020) 37:509–525
https://doi.org/10.1007/s00357-019-09324-6

A Modified k-Means Clustering Procedure for Obtaining a Cardinality-Constrained Centroid Matrix

Naoto Yamashita[1] · **Kohei Adachi**[1]

Published online: 16 July 2019

Abstract

k-means clustering is a well-known procedure for classifying multivariate observations. The resulting centroid matrix of clusters by variables is noted for interpreting which variables characterize clusters. However, between-clusters differences are not always clearly captured in the centroid matrix. We address this problem by proposing a new procedure for obtaining a centroid matrix, so that it has a number of exactly zero elements. This allows easy interpretation of the matrix, as we may focus on only the nonzero centroids. The development of an iterative algorithm for the constrained minimization is described. A cardinality selection procedure for identifying the optimal cardinality is presented, as well as a modified version of the proposed procedure, in which some restrictions are imposed on the positions of nonzero elements. The behaviors of our proposed procedure were evaluated in simulation studies and are illustrated with three real data examples, which demonstrate that the performances of the procedure is promising.

Keywords Clustering · Interpretation · Cardinality constraint · Alternating least squares

1 Introduction

k-means clustering is one of the most popular procedures for classifying the rows of an observations × variables data matrix into a small number of clusters (Aggarwal and Reddy 2013). The k-means clustering procedure is widely used for classification purposes and recent advances in its development can be found in Steinley (2006). Various extensions and related procedures of the k-means clustering exist, fuzzy versions (Miyamoto et al. 2008), probabilistic models (Bock 1996), and variable selection procedures (Brusco and Cradit 2001), for example. Applications of k-means clustering can be found in various fields of sciences, such as biology (Jetti et al. 2014), environmental science (Dalton et al. 2016), agricultural science (Hyland et al. 2016), engineering (Peng et al. 2013), applied psychology

✉ Naoto Yamashita
nyamashita@hus.osaka-u.ac.jp

[1] Graduate School of Human Sciences, Osaka University, 1-2, Yamadaoka, Suita City, Osaka, Japan

(Cortina and Wasti 2005; Kuerbis et al. 2014), and experimental psychology (Schloss et al. 2015; Alsius et al. 2016; Slobodenyuk et al. 2015).

For the n-observations × p-variable data matrix $\mathbf{X}$, k-means clustering is formulated as minimizing the least squares loss function

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

over $\mathbf{M}$ and $\mathbf{Y}$. Here, $\mathbf{M} = \{m_{il}\}(i = 1, \cdots, n; l = 1, \cdots, k)$ is an n-observations × k-clusters binary membership matrix and $\mathbf{Y} = \{y_{jl}\}(j = 1, \cdots, p)$ is a p-variables × k-cluster centroid matrix. $\mathbf{x}_{(i)}$ and $\mathbf{y}_l$ denote the i-th row vector and l-th column vector of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The operator $||\mathbf{X}||^2 = \mathrm{tr}(\mathbf{X}'\mathbf{X})$ denotes the sum of the squared elements in $\mathbf{X}$. The minimization of Eq. 1 can be attained by using the iterative algorithm proposed by MacQueen (1967), in which $\mathbf{M}$ and $\mathbf{Y}$ are alternately updated.

The resulting membership matrix $\mathbf{M}$ shows how observations are classified into clusters. In the interpretation of what variables characterize the clusters, the centroid matrix $\mathbf{Y}$ plays a key role. Let us illustrate such an interpretation process with Adachi's (2006) 13 (jobs) × 12 (attributes) data matrix $\mathbf{X}$ called job impression data in this article, which describes the extent to which the jobs are described by the impression attributes. We applied k-means clustering to the column-centered version of the data matrix, with the number of clusters set to 4. Table 1 shows the resulting centroid matrix. Its j-th row equals the average vector of the observations classified into the l-th cluster and represents the variables that characterize the l-th cluster. For example, the third cluster (C3) is characterized by the impressions of "powerful," "strong," and "fast," as the elements for them in C3 show larger values than those in the other clusters, where police, journalist, sailor, and athlete are classified into C3. The within-column and between-column contrasts in the centroids help to capture which variables feature clusters, although these contrasts are not always observed clearly in the centroid matrix.

A typical strategy to clarify the contrasts is to ignore the elements close to zeros (those less than 0.4 in absolute value, for example) in the centroid matrix and regard them as zeros. For example, we can consider that the first cluster (C1) is characterized by the "useful," "stubborn," and "busy" impression, by ignoring the elements having absolute values less than 0.4. This strategy is not desirable, however, because which elements can be

Table 1 Estimated centroid matrix by k-means with four clusters to job impression data

	C1	C2	C3	C4
Admirable	0.338	– 0.379	0.121	0.005
Useful	0.557	– 0.393	0.182	– 0.276
Good	0.205	– 0.129	0.021	– 0.062
Large	– 0.010	– 0.418	0.382	0.057
Powerful	– 0.071	– 0.796	1.029	– 0.238
Strong	0.031	–0.711	1.014	– 0.436
Fast	– 0.333	0.125	0.950	– 1.100
Noisy	– 0.279	0.246	0.871	– 1.212
Young	0.019	0.686	0.461	– 1.548
Honest	0.443	– 0.332	0.168	– 0.224
Stubborn	0.569	– 1.239	0.386	0.569
Busy	0.590	0.282	0.207	– 1.243

ignored depends on the users' decisions. Such decisions are both subjective and potentially erroneous, because, in fact, they weaken the initial fit achieved by the centroid matrix in an intuitive manner.

In this article, considering the above problem of the interpretability of the resulting centroid matrix, we propose a new clustering procedure that produces an easily interpreted centroid matrix. We call this procedure cardinality-constrained k-means clustering (CCKM). In CCKM, a number of the elements in the centroid matrix $\mathbf{Y}$ are constrained to be zeros, where $Card(\mathbf{Y})$ for the cardinality of $\mathbf{Y}$, i.e., its number of nonzero elements. The constraint is expressed as

$$Card(\mathbf{Y}) = c \tag{2}$$

with c a pre-specified integer. That is, our proposed CCKM is formulated as minimizing Eq. 1 subject to Eq. 2. Here, it should be noted that it is unknown which elements in $\mathbf{Y}$ are zero/nonzero. They are also estimated optimally in CCKM. This provides the centroid matrix with $pk - c$ zero elements, which facilitates the interpretation of $\mathbf{Y}$, as we may ignore the zero elements in $\mathbf{Y}$. Here, it is noteworthy that no subjective decision is involved in what elements are ignored; as described above, which elements are to be zeros is estimated optimally.

The classic work Gordon (1973) firstly introduced constrained clustering procedure, in which a priori information as to clusters are incorporated to clustering. Such a priori information is thoroughly discussed in DeSarbo and Mahajan (1984); some pairs of objects are constrained to be in the same or different cluster, for example. These constraints are helpful for obtaining interpretable and valid clusters. The related works can be found in Steinley and Hubert (2008) and Basu et al. (2008). These procedures require clearly defined information as to cluster structures although is often unavailable before applying clustering in the case of exploratory data analysis. The proposed procedure in this paper does not require such external information with combination of cardinality selection procedure introduced in Section 3.

The proposed procedure is also related to variable selection technique in clustering (Brusco and Cradit 2001). These procedures are designed to obtain appropriate cluster structure even if some musking variables (Fowlkes and Mallows 1983) exist, which often spoils clustering result. In order to accomplish this, variable selection procedure specifies the set of variables which manifest cluster structure by various strategies, which are detailed and compared in Steinley and Brusco (2008). CCKM differs from the variable selection procedures in that it is aimed to improve interpretability of cluster centroids and therefore, the set of variable corresponds to clusters are not always identical.

The remaining parts of this paper are organized as follows. In the following section, we present the algorithm for minimizing Eq. 1 under the cardinality constraint (2). In Section 3, a cardinality selection procedure is proposed, in which the best c is chosen with an information criterion. In Section 4, a modified version of CCKM is presented, in which cardinality constraints are imposed row/column-wise, although Eq. 2 is imposed matrix-wise. Two numerical simulations and three real data examples are presented in Sections 5 and 6, respectively, in order to assess and illustrate the performances of the proposed procedure. Section 7 is devoted to a general discussion.

1.1 Related Method: Factor Rotation

Interpretability of solutions is of importance also in the multivariate analysis procedures other than clustering. Factor rotation is a well-known technique to obtain interpretable solutions in factor analysis (FA), in which its initial solution is transformed into more interpretable one of simple structure (Browne 2001). It is based on the rotational indeterminacy of the FA solution, which allows a factor loading matrix be transformed in a post hoc manner. Varimax rotation (Kaiser 1974), Oblimin rotation (Harman 1976), and Promax rotation (Hendrickson and White 1964) are known as common procedures for factor rotation. More generally, rotation can be applied to some other multivariate analysis techniques, not limited for factor analysis; Yamashita (2012) and Satomura and Adachi (2013) proved that solutions can be rotated in canonical correlation analysis.

In clustering, however, such post hoc transformations of solutions are not feasible. We therefore propose CCKM for improving the interpretability of solutions in clustering, and it is surely beneficial for practical users. Also, as a special case of CCKM, we propose RC-CCKM to produce the centroid matrix having a simple structure, by means of restricting the row/column cardinality of the centroid matrix. Such a structure is desired also in FA (Thurstone 1947; Ullman 2006).

2 Algorithm

The optimization algorithm for CCKM is outlined in Section 2.1. It is composed of two steps called the M- and Y-steps, which are iteratively alternated until convergence is reached. These steps are described in detail in Section 2.2.

2.1 Outline: Whole Algorithm

Our proposed CCKM is formulated as

$$minimize\ f(\mathbf{M}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2\ over\ \mathbf{M}\ and\ \mathbf{Y} \tag{3}$$

subject to the cardinality constraint (2) and the membership constraint imposed on **M** such that

$$m_{il} \in \{0, 1\}\ and\ \sum_{l} m_{il} = 1. \tag{4}$$

Two parameter matrices are alternately updated in the M- and Y-steps, respectively, starting from multiple sets of initial values in order to avoid accepting a local minimum as a final solution. In each step, Eq. 1 is minimized over **M** or **Y** with the other parameter matrix kept fixed. The CCKM algorithm is thus summarized as follows.

Step 1. Set $t = 0$.
Step 2. Set initial values for **M** and **Y**.
Step 3. (M-step) Update **M** to that which minimizes $f(\mathbf{M}, \mathbf{Y})$ subject to Eq. 4 with **Y** fixed.
Step 4. (Y-step) Update **Y** to that which minimizes $f(\mathbf{M}, \mathbf{Y})$ subject to Eq. 2 with **M** fixed.
Step 5. If the current **M** has an empty column filled with zeros, return to Step 2.
Step 6. Increase t by one and go to Step 7, if the decrease in the Eq. 1 value from the previous round is less than 1.0×10^{-7}; otherwise, return to Step 3.
Step 7. Update $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}$ by the current **M** and **Y** if $f(\hat{\mathbf{M}}, \hat{\mathbf{Y}}) > f(\mathbf{M}, \mathbf{Y})$ or $t = 1$.
Step 8. If $t = t_{max}$, accept $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}$ as the final solution; otherwise, return to Step 2.

The purpose of Step 5 is to avoid a solution with an empty cluster to which no observations belong. In order to avoid accepting a local minimum, Steps 2 to 6 are repeated t_{max} times starting from different initial values. Among the resulting multiple solutions, that with the lowest (1) value is selected as the optimal solution. The update formulas used in the M- and Y-steps are presented in the following section.

2.2 Optimization in M-Step and Y-Step

For the M-step, the corresponding step in the k-means clustering (MacQueen 1967) can be used: for fixed $\mathbf{Y}$, the optimal $\mathbf{M} = \{m_{il}\}$ minimizing Eq. 1 is given by

$$m_{il} = \begin{cases} 1 & (l = \underset{l}{\arg\min}\, f(\mathbf{M}, \mathbf{Y})) \\ 0 & (otherwise) \end{cases}, \tag{5}$$

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

The problem in the Y-step cannot be solved straightforwardly and we therefore need a trick. As such a one, we use the fact that Eq. 1 can be decomposed as

$$f(\mathbf{M}, \mathbf{Y}) = ||\mathbf{X} - \mathbf{M}\mathbf{A}'||^2 + ||\mathbf{D}^{1/2}(\mathbf{A} - \mathbf{Y})||^2. \tag{6}$$

Here, $\mathbf{D} = \text{diag}\{d_{11}, \cdots, d_{ll}, \cdots, d_{kk}\}$ denotes the $k \times k$ diagonal matrix with d_{ll} the number of the observations classified into the l-th cluster ($l = 1, \cdots, k$), while $\mathbf{A}$ is defined as

$$\mathbf{A} = \mathbf{X}'\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}, \tag{7}$$

the l-th column of which contains the averages of the observation in the l-th cluster. The identity in Eq. 6 can be proved as follows. Equation 1 can be rewritten as

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

with its last term vanishing as

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

In the right-hand side of Eq. 6, only $g(\mathbf{Y}) = ||\mathbf{D}^{1/2}(\mathbf{A} - \mathbf{Y})||^2$ is relevant to $\mathbf{Y}$. Thus, our task is to minimize $g(\mathbf{Y})$ subject to Eq. 2. The minimum can be found using the fact that $g(\mathbf{Y})$ satisfies the following identity and inequality:

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

where Z denotes the set of $pk - c$ pairs (j, l)s for the elements y_{jl}s to be zero, while $Z^{\perp}$ is the complement of Z. The inequality in Eq. 10 shows that $g(\mathbf{Y})$ attains its lower limit $\sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2$ when the elements y_{jl}s in the second term in Eq. 10 is equal to zero, that is, when the y_{jl}s with $(j, l) \in Z^{\perp}$ is set equal to zero. Further, the limit $\sum_{(j,l)\in Z} d_{ll}^{1/2} a_{jl}^2$ is minimal when Z contains the indices for the $pk - c$ smallest a_{jl}^2s among all squared elements in $\mathbf{A}$. Therefore, the optimal $\mathbf{Y}$ is obtained as

$$y_{jl} = \begin{cases} 0 & (\textit{iff}\ a_{jl}^2 \leq a_{[pk-c]}^2) \\ a_{jl} & (otherwise) \end{cases} \tag{11}$$

for $l = 1, \cdots, k$ and $j = 1, \cdots, p$, where $a_{[pk-c]}^2$ denotes the $(pk - c)$-th smallest value among all $a_{jl}^2 s$.

Equations 5 and 11 are used for the updates in M- and Y-steps, respectively. They guarantee the monotonic decrement in the the $f(\mathbf{M}, \mathbf{Y})$ value. In the following simulation studies and real data examples, we used 300 different initial values for **M** and **Y**; i.e., $t_{max} = 300$.

3 Cardinality Selection Based on Information Criteria

In the CCKM algorithm, the cardinality of centroid matrix **Y** has to be set to a positive integer c, as in Eq. 2. In this article, the minimum and maximum of c, $c_{min},\ c_{max}$, are defined as

$$c_{min} = p, \quad c_{max} = pk. \tag{12}$$

It should be noted that **Y** has p non-zero elements when **Y** has a perfect cluster structure; i.e., each variable is associated with only one cluster. The selection of the number of non-zero elements in **Y** can be viewed as a model selection problem, since the selection partially specifies the model part of CCKM, $\mathbf{MY}'$, fitted to the given data matrix **X**. Thus, information criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), for the model selection problem can be suitable for determining c, which directly constrains the cardinality of the model fitted to the dataset. In this study, we propose using AIC and BIC to select the "best" c among the interval $[c_{min}, c_{max}]$.

Let $\mathbf{E} = \{e_{ij}\}$ be the matrix of errors defined as $\mathbf{E} = \mathbf{X} - \mathbf{MY}'$ and assume that data matrix **X** is modeled as $\mathbf{X} = \mathbf{MY}' + \mathbf{E}$ with e_{ij} distributed independently and identically according to $N(0, \sigma^2)$ for all i and j. Here, $N(0, \sigma^2)$ represents normal distribution with its mean zero and variance σ^2. Then, it can be shown that the least squares and maximum likelihood estimates for CCKM are equivalent. Under the above assumption, the log-likelihood function to be maximized in ML estimation is expressed as

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

Its maximization is equivalent to minimizing least squares function (1). For a positive integer c, the maximum of $l(\mathbf{M}, \mathbf{Y})$ is attained as

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

where $f_{min}(c)$ denotes the attained function value of Eq. 1. Using Eq. 14, the information criteria $AIC(c)$ and $BIC(c)$ for a specific c are defined as

$$AIC(c) = np \times \log f_{min}(c) + 2\nu(c), \tag{15}$$

$$BIC(c) = np \times \log f_{min}(c) + \log(np) \times \nu(c), \tag{16}$$

where $\nu(c)$ denotes the number of parameters to be estimated and equals the sum of the numbers of the memberships in **M**, the nonzero elements in **Y**, and error variance σ^2:

$$\nu(c) = n + c + 1. \tag{17}$$

Therefore, the best c can be given by $c = \underset{c_{min} \leq c \leq c_{max}}{\arg\min}\ AIC(c)$ or $BIC(c)$. AIC and BIC were originally proposed as model selection criteria and also used for cardinality selection in studies in the literature Adachi and Trendafilov (2015, 2017). This approach is not considered to be computationally efficient, however, because CCKM involves the optimization with respect to a binary membership matrix, as in the standard k-means, and therefore is sensitive to local minima. To select c with $AIC(c)$ or $BIC(c)$ is not feasible because a heavy computational load is required for all runs, especially when the data matrix is large (i.e., **X** contains many objects and variables). In such a situation, the centroid matrix to be estimated is also large, although CCKM must facilitate the interpretation of such a large matrix.

In order to find a suitable c value with a 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 $c_{current} = c_{min}$ and $S = c_{max} \times S_{initial}$.

Step 2. Repeat Steps 2 and 3 for $S > 1$.

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

(a) Set $c = c_{current}$ and compute

$$\Delta AIC(c) = AIC(c+1) - AIC(c) \tag{18}$$

or

$$\Delta BIC(c) = BIC(c+1) - BIC(c). \tag{19}$$

(b) If $\Delta AIC(c)$ or $\Delta BIC(c)$ is smaller than 0, set $c_{current} = c_{current} + S$ and return to Step 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 $c = c_{current}$ and compute $\Delta AIC(c)$ or $\Delta BIC(c)$.

(b) If $\Delta AIC(c)$ or $\Delta BIC(c)$ is greater than 0, set $c_{current} = c_{current} - S$ and return to Step 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 $\Delta AIC(c)$ or $\Delta BIC(c)$ becomes positive; otherwise, repeat *forward search* until $\Delta AIC(c)$ or $\Delta BIC(c)$ becomes negative.

The above algorithm seeks c that minimizes $AIC(c)$ or $BIC(c)$ within the range $[m_{min}, m_{max}]$ by repeating the forward and backward searches and reducing the step size S through iteration. The rate of decrement in the step size is controlled by $S_{decrease}$ and the initial step size is defined as $c_{max} \times S_{initial}$. The total computational cost is therefore dramatically reduced in comparison with that incurred by performing CCKM with computation of $AIC(c)$ or $BIC(c)$ for all cs. In the following simulation and real data examples, we set $S_{initial} = 0.9$ and $S_{decrease} = 0.7$, settings that were empirically confirmed to behave well.

4 CCKM with Row/Column-Wise Cardinality Constraint (RC-CCKM)

While a matrix-wise cardinality is constrained by cardinality parameter c in CCKM, we can also consider its modified version subject to row- and column-wise cardinality constraints. By combining row- and column-wise constraints, we can restrict both the cardinality of the centroid matrix and the positions of the nonzero elements, so that the resulting matrix approximates a simple structure. The utility of this approach can be illustrated with an example of a 5 variables × 3 clusters centroid matrix that has several zero elements, but is not easily interpretable:

$$\mathbf{Y} = \begin{bmatrix} & & * \\ * & * & * \\ & & * \\ * & * & * \\ & & * \end{bmatrix}. \tag{20}$$

Here, the cardinality is 10 and $*$ represents a nonzero element. Although this matrix has more zero elements than unconstrained ones, it is still difficult to interpret in that the centroid matrix has a column (the third column) filled with non-zero elements, which indicates that all the variables are associated with the third cluster. To interpret of such a cluster, abstraction and integration of all variables are required, and it is not always straightforward to name the cluster. Similarly, the second and fourth rows are filled with non-zeros. We therefore prefer the centroid matrix

$$\mathbf{Y} = \begin{bmatrix} * & * & \\ * & * & \\ & * & * \\ * & & * \\ & * & * \end{bmatrix} \tag{21}$$

because it does not contain any row/column vector filled with non-zeros in spite of its cardinality equaling that of Eq. 20. In order to obtain a centroid matrix as in Eq. 21, the positions of nonzero elements, in other words, the cardinality of rows and columns, have to be restricted. Such constraints are defined as

$$Card(\mathbf{y}_{(j)}) = r(j), \ Card(\mathbf{y}_l) = c(l) \tag{22}$$

where $r(j)$ and $c(l)$ denote the cardinality of the row and column of $\mathbf{Y}$, respectively.

To find the matrices $\mathbf{M}$ and $\mathbf{Y}$ that minimize Eq. 1 under the above constraint, the Y-step in Section 2.2 can be modified as follows. To minimize Eq. 1 subject to the constraint $Card(\mathbf{y}_{(j)}) = r(j)$, the set Z is redefined as

$$Z = \{(j,l) | a_{jl}^2 \leq a_{j[r(j)]}^2\}, \tag{23}$$

where $\{a_{j[k-r(j)]}^2\}$ denotes the $\{k - r(j)\}$-th smallest element among $a_{j1}^2, \cdots, a_{jk}^2$ for $j = 1, \cdots, p$. Each row is therefore updated by using the above Z and referencing the squared elements of $\mathbf{A}$ in Eq. 7. In a parallel manner, under $Card(\mathbf{y}_l) = c(l)$, Z is redefined as

$$Z = \{(j,l) | a_{jl}^2 \leq a_{[c(l)]l}^2\} \tag{24}$$

with $a_{[p-c(l)]l}^2$ the $\{p - c(l)\}$-th smallest element among $a_{1l}^2, \cdots, a_{pl}^2$ for $l = 1, \cdots, k$.

We refer to the above procedure as CCKM with row/column-cardinality constraint (RC-CCKM). The performance of the procedure is demonstrated later in one of the real data examples.

5 Simulation Studies

In order to assess the behaviors of the CCKM algorithm presented in Section 2 and the AIC/BIC-based cardinality selection procedure presented in Section 3, we performed two numerical simulation studies. The behaviors to be assessed are the following two. (1) The correctness of the identification of the true cardinality is identified by the cardinality selection procedure and (2) the recovery by CCKM of the true parameters from which artificial datasets are synthesized. Therefore, the purpose of the first simulation study was to assess the accuracy of the cardinality selection, and of the second to evaluate the performance of the parameter recovery.

5.1 Accuracy of Cardinality Selection

First, we examined the accuracy of true cardinality in the cardinality selection procedure. A hundred data matrices $\mathbf{X}$ of $n = 100$ by $p = 30$ were randomly generated with setting $k = 3$, as follows.

Step 1. A positive integer c_T was randomly drawn from the interval $[0.1 \times pk, 0.9 \times pk]$ and used for the true cardinality of a centroid matrix.

Step 2. The c_T nonzero elements in $\mathbf{Y}$ were drawn from the uniform distribution $U(1, 5)$, with their positions and signs randomly chosen.

Step 3. The true membership matrix $\mathbf{M}$ was formed by randomly assigning n observations to k clusters.

Step 4. The elements of $n \times p$ error matrix $\mathbf{E}$ were drawn from the standard normal distribution $N(0, 1)$.

Step 5. Data matrix $\mathbf{X}$ of n-observations $\times$ p-variables was synthesized with

$$\mathbf{X} = \mathbf{M}\mathbf{Y}' + \mu(\rho)\mathbf{E} \tag{25}$$

where $\mu(\rho)$ is defined as

$$\mu(\rho) = \sqrt{\frac{1-\rho}{\rho} \times \frac{||\mathbf{M}\mathbf{Y}'||^2}{||\mathbf{E}||^2}} \tag{26}$$

with ρ being the rate of the variance explained.

In Step 5, $\mu(\rho)$ was used for controlling the level of errors, so that ρ approximated the proportion of the variances in $||\mathbf{X}||^2$ accounted for by the model part $\mathbf{M}\mathbf{Y}'$ (Adachi 2009; Yamashita and Mayekawa 2015). We considered the medium error level with $\rho = 0.70$, which represents the 70% of the variance of $\mathbf{X}$ explained by $\mathbf{M}\mathbf{Y}'$. For each of 100 $\mathbf{X}$s, the best cardinality was selected with AIC and BIC: c_{AIC} and c_{BIC} were used for the selected cardinality, respectively. The bias of the estimated cardinality was evaluated with the standardized difference (SD) defined as

$$SD(m) = (pk)^{-1}(c_T - c). \tag{27}$$

Figure 1 shows the boxplots of the resulting $SD(c_{AIC})$ and $SD(c_{BIC})$ in the medium-error condition. The SD values are found to approximate zero in the BIC-based cardinality selection, which indicates that the BIC-based procedure almost perfectly specifies the true cardinality, although it slightly overestimates it in a few cases. Conversely, the AIC-based selection tended to considerably underestimate the true cardinality, which is seen in the result that the discrepancy between the true and estimated cardinality is over 7% of all entries in $\mathbf{Y}$ at 50 percentile of $SD(c)$. These results show that the BIC-based is more precise than the AIC-based selection, suggesting that the former should be used.

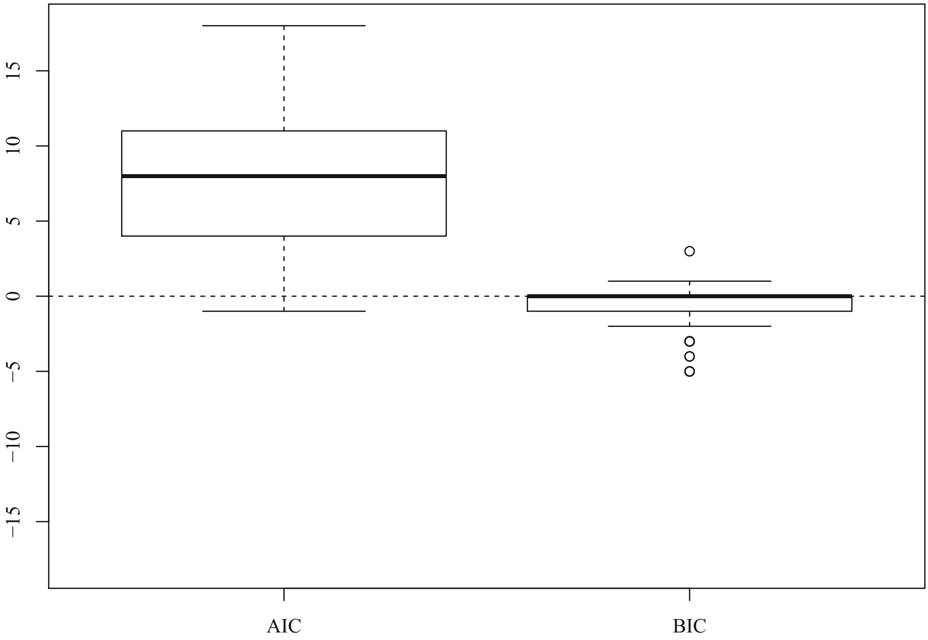

Fig. 1 Boxplots of $SD(c_{AIC})$ and $SD(c_{BIC})$ as indices of discrepancy between true and estimated cardinality by AIC and BIC criterion

5.2 Accuracy of Parameter Estimation

In order to assess how well the parameter matrices are recovered in the CCKM algorithm, we considered the cases where $(n, p) = (100, 30)$ and $(30, 100)$ and three error levels $(\rho = 0.90, 0.70, 0.50)$. A hundred artificial datasets $\mathbf{X}$ were generated as described in the previous section. Next, CCKM was applied to $\mathbf{X}$ with its cardinality set as $c = c_T$ (identical to the true cardinality), $c = c_T - 0.05 \times pk$ (5% fewer than the true cardinality), and $c = c_T + 0.05 \times pk$ (5% more than the true cardinality) in order to assess the sensitivity to cardinality misidentification. For each of 2 (dimension of data matrix; $(n, p) = (100, 30), (30, 100)) \times 3$ (error level; $\rho = 0.90, 0.70, 0.50) \times 3$ (cardinality setting; 5% fewer, identical to true cardinality, and 5% more) combinations, we generated 100 datasets in the manner described in the previous section. The resulting parameters $(\hat{\mathbf{M}}, \hat{\mathbf{Y}})$ were compared to their true counterparts and the accuracy of their recovery was evaluated by the following indices. For the membership matrix, we compared $\hat{\mathbf{M}}$ and $\mathbf{M}$ in terms of the Adjusted Rand Index (ARI) (Rand 1971; Hubert and Arabie 1985) of these matrices, ranged from 0 to 1 with $ARI = 1$ representing the perfect coherence of the two partitions shown by $\hat{\mathbf{M}}$ and $\mathbf{M}$. The proximity between $\hat{\mathbf{Y}} = \{\hat{y}_{jl}\}$ and its true counterpart $\mathbf{Y} = \{y_{jl}\}$ was evaluated with Averaged Absolute Error (AAE) defined as $AAE(\hat{\mathbf{Y}}, \mathbf{Y}) = (pk)^{-1} \sum_{j,l} |\hat{y}_{jl} - y_{jl}|$, which indicates the averaged discrepancy between the pk elements in $\hat{\mathbf{Y}}$ and $\mathbf{Y}$. It should be noted that, before computing the four indices, we must choose the $k \times k$ permutation matrix that minimizes $||\hat{\mathbf{Y}}\mathbf{P} - \mathbf{Y}||^2$, to eliminate the freedom of row-permutation of $\hat{\mathbf{Y}}$ shown as $||\mathbf{X} - \mathbf{M}\mathbf{Y}'||^2 = ||\mathbf{X} - (\mathbf{M}\mathbf{P})(\mathbf{Y}\mathbf{P})'||$. Therefore, the permutation of $\mathbf{P}$ was chosen that minimizes $||\hat{\mathbf{Y}}\mathbf{P} - \mathbf{Y}||^2$ among all possible $p!$ permutation matrices. We use $\hat{\mathbf{M}}\mathbf{P}$ for $\hat{\mathbf{M}}$ and $\hat{\mathbf{Y}}\mathbf{P}$ for $\hat{\mathbf{Y}}$ with the chosen $\mathbf{P}$.

The CCKM algorithm is run starting from t_{max} different initial values in order to avoid accepting a local minimum as the final solution. In this simulation, we ran the CCKM algorithm from 300 different initial values; i.e., $t_{max} = 300$.

Figures 2 and 3 show the parameter recovery results in the two cases of data sizes. Overall, the parameters were correctly recovered by CCKM, in that the ARI indices attained their maximum 1 in almost all cases, and that the discrepancy between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ indicated by the AAE values are sufficiently small. Even in the case where $\rho = 0.50$, where the error variance amounts to the half of the total variance of the dataset, the AAE value is lower than 0.3 at 50 percentile. Further, even if the cardinality is misidentified, the levels of the ARI and AAE values are still satisfactory. Thus, it can be considered that the true cardinality can be identified fairly accurately in CCKM. These results allow us to conclude that the performances of CCKM is suitable for dealing with practical problems, in that the CCKM almost perfectly recovers the true parameters.

6 Real Data Examples

In this section, we illustrate CCKM with the cardinality selection procedure, using three real data examples. Further, the modified version of CCKM with row/column-cardinality constraints are finally illustrated with an additional example.

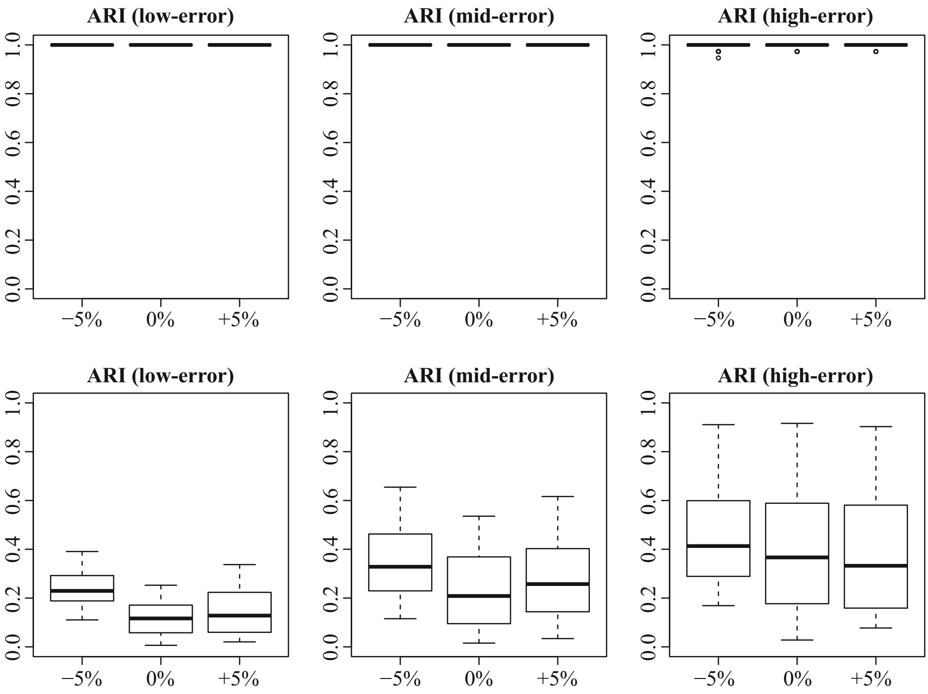

Fig. 2 Boxplot of ARI (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (100, 30)$

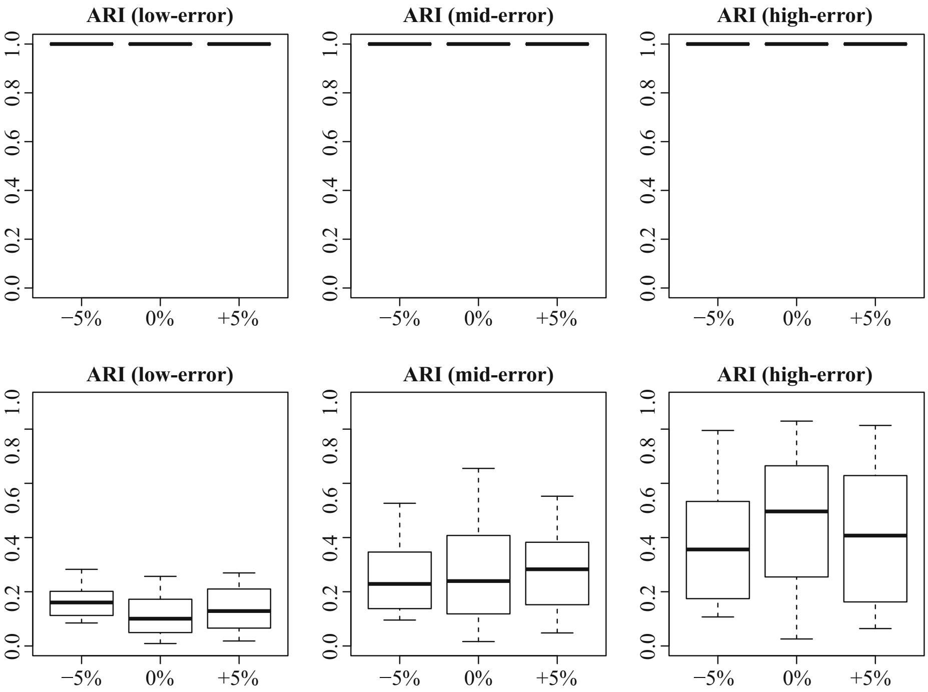

Fig. 3 Boxplot of ARI (adjusted Rand index) and $AAE(\hat{\mathbf{Y}}, \mathbf{Y})$ values in the case with $(n, p) = (30, 100)$

6.1 Example 1: Fisher's Iris Data

CCKM was applied to Fisher's (1936) Iris data, in which the 150 observations sampled from 3 species were measured with respect to 4 variables. Note that the data matrix was column-wise standardized beforehand. In order to find the optimal cardinality, the BIC-based cardinality selection procedure was used and the results suggested that $c = 8$ is the best. We also applied k-means clustering to the dataset for comparison. The estimated centroid matrices are shown in Table 2.

In the centroid matrix estimated by CCKM in Table 2, we can find that the first cluster is contrasted to the second one for Sepal.Length and Sepal.Width. Further, the second cluster is different from the other clusters for Sepal.Width, with Versicolor in the second cluster

Table 2 Estimated centroid matrices by CCKM with $c = 10$ (left) and k-means (right); a blank cell shows exact-zero element

	CCKM			k-means		
	C1	C2	C3	C1	C2	C3
Sepal.Length	1.030		− 0.999	1.132	− 0.050	− 1.011
Sepal.Width		− 0.969	0.903	0.088	− 0.880	0.850
Petal.Length	0.940		− 1.299	0.993	0.347	− 1.301
Petal.Width	0.969		− 1.252	1.014	0.281	− 1.251

Table 3 Contingency table for species vs. the estimated partitions by CCKM and k-means

	CCKM($c = 8$)			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

characterized by narrow sepals. The contrast between clusters is clearer in the centroid matrix of CCKM than in that of the resulting k-means clustering, because the former has exactly zero elements. Further, as shown in Table 3, the three clusters are associated with the three species, Setosa, Versicolor, and Virginica, respectively. It can be seen that the estimated memberships correspond to the species, since $(49 + 37 + 42)/150 = 85.3\%$ of 150 observations are correctly classified into the true species, while $(50 + 39 + 36)/150 = 83.3\%$ in the k-means clustering. The ARI (with its 95 % confidence interval by Steinley et al. (2016)) between the partition obtained by CCKM for the three species was 0.645 ([0.626, 0.664]), while $ARI = 0.620$ ([0.601, 0.638]) for k-means clustering. Also, ARI for the two partitions was 0.832 ([0.813, 0.851]), and it suggests that two partitions obtained by CCKM and k-means are similar, even if the cardinality of centroid matrix is restricted in CCKM. These results demonstrate that CCKM yields an easily interpreted centroid matrix subject to the cardinality constraint, and the classification accuracy is improved.

6.2 Example 2: Wine Data

The second example is that of the Wine data available at the UCI machine learning repository, in which the 178 types of wine are evaluated in terms of 13 chemical features. Note that the data matrix was column-wise standardized beforehand. CCKM was performed for the dataset with $c = 35$ selected by the BIC-based procedure. Table 4 shows the estimated

Table 4 Centroid matrix obtained by CCKM with cardinality of $r = 35$. The exact-zero elements are shown as blank cells

	C1	C2	C4	C4	C5
Alcohol	0.930	– 0.613	– 0.993	– 0.737	
Malic acid				– 0.843	0.930
Ash		1.960		– 1.225	
Alcalinity of ash	– 0.780	1.499	0.668	– 0.587	0.524
Magnesium	0.505	1.548	−0.701		
Total phenols	0.881				– 1.004
Flavonoids	0.958	0.491			– 1.233
Nonflavonoid phenols	– 0.611				0.743
Proanthocyanins	0.599				– 0.774
Color intensity			– 1.007	– 0.709	0.953
Hue				0.975	– 1.180
OD280 315	0.772				−1.275
Proline	1.204		– 0.807	– 0.652	

centroid matrix. It can be seen that each cluster is characterized by a small number of chemical features, which highlights the contrasts among clusters. For example, the contrast between the fourth and fifth clusters can be found in "Malic acid" and "Hue"; the observations classified into the fourth cluster are characterized by lower "Malic acid" and higher "Hue." Similarly, the other contrast between clusters can be found in several variables, such as "Ash," "Total phenols," and "Nonflavonoid phenols," which are clearly highlighted by the cardinality constrained estimation of nonzero and exact-zero elements.

Originally, the wines in the dataset are classified into three categories. In Table 5, we compare these categories and the clusters obtained by CCKM and k-means. Note that the number of clusters in k-means was set at 5 for comparison. It is interesting that clusters obtained by the two procedures are fairly similar, as indicated that $ARI = 0.715$ with its 95% interval $[0.700,\ 0.731]$ for the two partitions. For both procedures, the first and fifth clusters correspond to Category 1 and 3, respectively. Further, the wines in Category 2 is splitted into three clusters (second, third, and fourth clusters). It suggests that even if CCKM reduces cardinality of centroid matrix, it produces similar cluster structure to k-means clustering.

6.3 Example 3: Job Impression Data

Here, we illustrate RC-CCKM by using the job impression dataset that was also used in Section 1. Note that the data matrix was column-wise centered beforehand. Row- and column-wise cardinalities were set to $r(j) = k/2 = 2$ and $c(l) = p/2 = 6$ for $j = 1, \cdots, 12, l = 1, \cdots, 4$. This implies "Each row should contain at least one zero and each column should contain at least k zeros," which is a property of an easily interpreted matrix (Thurstone 1947).

The estimated centroid matrix is presented in Table 6. The resulting cardinality of the centroid matrix is 19, which is $19/48 \approx 39.6\%$ of all entries. As in the previous two examples, we can see clear differences between clusters and the homogeneity in each cluster. As the row-wise cardinality was set at 2, each row has at least two zero elements and the remaining nonzero elements show the contrast of extreme positive and negative values, as found in the rows of "powerful" and "strong," for example. Furthermore, each cluster corresponds to fewer variables, as compared with the results of k-means in Table 1, which facilitates the interpretation of clusters. It should be noted that the proportion of the variance explained by the CCKM is 60.8%, which is not substantially lower than the proportion of 68.9% in the k-means solution. Further, the memberships estimated in CCKM perfectly correspond to those in the k-means clustering. These results demonstrate promising performances of RC-CCKM: it gives easily interpreted solutions while the goodness of fit to data is retained.

Table 5 Contingency table for wine categories vs. the estimated partitions by CCKM and k-means

	CCKM($c = 35$)					k-means				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Category 1	56	2	0	1	0	54	4	0	1	0
Category 2	1	4	37	27	2	0	18	25	27	1
Category 3	0	1	0	0	47	0	0	1	0	47

Table 6 Estimated centroid matrix by CCKM with row/column-cardinality constraint and four clusters for job impression data; exact-zero elements are shown as blank cells

	C1	C2	C3	C4
Admirable	0.338			
Useful	0.557	– 0.393		
Good				
Large		– 0.418		
Powerful		– 0.796	1.029	
Strong		– 0.711	1.014	
Fast			0.950	– 1.100
Noisy			0.871	– 1.212
Young		0.686		– 1.548
Honest	0.443			
Stubborn	0.569	– 1.239		
Busy	0.590			– 1.243

7 Concluding Remarks

In this paper, we addressed the difficulty in interpreting the centroid matrix resulting from the standard k-means clustering. We proposed a new procedure called CCKM in which the cardinality of the centroid matrix is directly constrained ti improve its interpretability. CCKM produces a centroid matrix with reduced cardinality and its interpretation is easier than that of the standard k-means clustering, because between-cluster contrasts are highlighted by exact-zero elements. We also proposed a cardinality selection procedure and modified version of CCKM. The results of the simulation studies show that the BIC-based cardinality selection is more accurate than the AIC-based one, and the parameter estimation of CCKM is not sensitive to error contamination and misidentification of cardinality. Real data examples were presented to demonstrate the promising performance of CCKM and its modified version.

In clustering, the interpretability of solutions is of importance, as well as the classification accuracy. Thus, the cardinality constraint is considered to be useful for users in that the number of non-zero elements is directly associated with interpretability. Further, users can control the balance of low cardinality and model fit by tuning the cardinality parameter c within a restricted range. Therefore, we can conclude that the proposed procedure is suitable for extracting interpretable clusters.

Acknowledgments The authors are deeply grateful for the helpful comments and suggestions of the associate editor.

References

Adachi, K. (2009). Joint Procrustes analysis for simultaneous nonsingular transformation of component score and loading matrices. *Psychometrika*, *74*, 667–683.

Adachi, K., & Trendafilov, N.T. (2015). Sparse principal component analysis subject to prespecified cardinality of loadings. *Computational Statistics*, *31*, 1–25.

Adachi, K. (2006). *Multivariate data analysis*. Tokyo: Nakanishiya Shuppan. in Japanese.

Adachi, K., & Trendafilov, N.T. (2017). Sparsest factor analysis for clustering variables: a matrix decomposition approach. *Advances in Data Analysis and Classification*, *25*, 1–29.

Aggarwal, C.C., & Reddy, C.K. (2013). *Data clustering: algorithms and applications*. Boca Raton: CRC Press.

Alsius, A., Wayne, R.V., Paré, M., Munhall, K.G. (2016). High visual resolution matters in audiovisual speech perception, but only for some, Attention. *Perception, & Psychophysics*, *78*, 1472–1487.

Basu, S., Davidson, I., Wagstaff, K. (Eds.) (2008). *Constrained clustering: advances in algorithms, theory, and applications*. Boca Raton: CRC Press.

Bock, H.H. (1996). Probabilistic models in cluster analysis. *Computational Statistics & Data Analysis*, *23*, 5–28.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Brusco, M.J., & Cradit, J.D. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika*, *66*, 249–270.

Cortina, L.M., & Wasti, S.A. (2005). Profiles in coping: responses to sexual harassment across persons, organizations, and cultures. *Journal of Applied Psychology*, *90*, 182–192.

Dalton, C., Jennings, E., O'dwyer, B., Taylor, D. (2016). Integrating observed, inferred and simulated data to illuminate environmental change: a limnological case study. *Biology and Environment: Proceedings of the Royal Irish Academy*, *116*, 279–294.

DeSarbo, W.S., & Mahajan, V. (1984). Constrained classification: the use of a priori information in cluster analysis. *Psychometrika*, *49*, 187–215.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, *7*, 179–188.

Fowlkes, E.B., & Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*, 553–569.

Gordon, A.D. (1973). 359. Note: Classification in the presence of constraints. *Biometrics*, *29*, 821–827.

Harman, H.H. (1976). *Modern factor analysis*, 3rd edn. Chicago: University of Chicago Press.

Hendrickson, A.E., & White, P.O. (1964). PROMAX: a quick method for rotation to oblique simple structure. *British Journal of Mathematical and Statistical Psychology*, *17*, 65–70.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*, 193–218.

Hyland, J.J., Jones, D.L., Parkhill, K.A., Barnes, A.P., Williams, A.P. (2016). Farmers' perceptions of climate change: identifying types. *Agriculture and Human Values*, *33*, 323–339.

Jetti, S.K., Vendrell-Llopis, N., Yaksi, E. (2014). Spontaneous activity governs olfactory representations in spatially organized habenular microcircuits. *Current Biology*, *24*, 434–439.

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika*, *39*, 31–36.

Kuerbis, A., Armeli, S., Muench, F., Morgenstern, J. (2014). Profiles of confidence and commitment to change as predictors of moderated drinking: a person-centered approach. *Psychology of Addictive Behaviors*, *28*, 1065–1076.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 281–297.

Miyamoto, S., Ichihashi, H., Honda, K. (2008). *Algorithms for fuzzy clustering*. Berlin: Springer.

Peng, X., Zhou, C., Hepburn, D.M. (2013). Application of K-Means method to pattern recognition in on-line cable partial discharge monitoring. *IEEE Transactions on Dielectrics and Electrical Insulation*, *20*, 754–761.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Satomura, H., & Adachi, K. (2013). Oblique rotation in canonical correlation analysis reformulated as maximizing the generalized coefficient of determination. *Psychometrika*, *78*, 526–573.

Schloss, K.B., Hawthorne-Madell, D., Palmer, S.E. (2015). Ecological influences on individual differences in color preference, Attention. *Perception, & Psychophysics*, *77*, 2803–2816.

Slobodenyuk, N., Jraissati, Y., Kanso, A., Ghanem, L., Elhajj, I. (2015). Cross-modal associations between color and haptics, Attention. *Perception, & Psychophysics*, *77*, 1379–1395.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*, 1–34.

Steinley, D., & Brusco, M.J. (2008). Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika*, *73*, 125–144.

Steinley, D., Brusco, M.J., Hubert, L. (2016). The variance of the adjusted Rand index. *Psychological Methods*, *21*, 261–272.

Steinley, D., & Hubert, L. (2008). Order-constrained solutions in K-means clustering: even better than being globally optimal. *Psychometrika*, *73*, 647–664.

Thurstone, L.L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Ullman, J.B. (2006). Structural equation modeling: reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*, 33–50.

Yamashita, N. (2012). Canonical correlation analysis formulated as maximizing sum of squared correlations and rotation of structure matrices. *The Japanese Journal of Behaviormetrics*, *39*, 1–9. (in Japanese).

Yamashita, N., & Mayekawa, S. (2015). A new biplot procedure with joint classification of objects and variables by fuzzy c-means clustering. *Advances in Data Analysis and Classification*, *9*, 243—266.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.