Bias-correction fuzzy clustering algorithms

Miin-Shen Yang a,b, Yi-Cheng Tian a,b

a Department of Applied Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan
b Center for General Education, Hsin Sheng College of Medical Care and Management, Longtan, Taiwan

Abstract

Fuzzy clustering is generally an extension of hard clustering and it is based on fuzzy membership partitions. In fuzzy clustering, the fuzzy c-means (FCM) algorithm is the most commonly used clustering method. Numerous studies have presented various generalizations of the FCM algorithm. However, the FCM algorithm and its generalizations are usually affected by initializations. In this paper, we propose a bias-correction term with an updating equation to adjust the effects of initializations on fuzzy clustering algorithms. We first propose the so-called bias-correction fuzzy clustering of the generalized FCM algorithm. We then construct the bias-correction FCM, bias-correction Gustafson and Kessel clustering and bias-correction inter-cluster separation algorithms. We compared the proposed bias-correction fuzzy clustering algorithms with other fuzzy clustering algorithms by using numerical examples. We also applied the bias-correction fuzzy clustering algorithms to real data sets. The results indicated the superiority and effectiveness of the proposed bias-correction fuzzy clustering methods.

1. Introduction

Clustering is a method for determining the cluster structure of a data set such that objects within the same cluster demonstrate maximum similarity and objects within different clusters demonstrate maximum dissimilarity. Numerous clustering theories and methods have been evaluated in the literature (see Jain and Dubes [10] and Kaufman and Rousseeuw [11]). In general, the most well-known approaches are partitional clustering methods based on an objective function of similarity or dissimilarity measures. In partitional clustering methods, the k-means (see MacQueen [14] and Pollard [20]), fuzzy c-means (FCM) (see Bezdek [2] and Yang [23]), and possibilistic c-means (PCM) algorithms (see Krishnapuram and Keller [12], Honda et al. [8], and Yang and Lai [24]) are the most commonly used approaches.

Fuzzy clustering has received considerable attention in the clustering literature. In fuzzy clustering, the FCM algorithm is the most well-known clustering algorithm. Previous studies have proposed numerous extensions of FCM clustering (see Gath and Geva [4], Gustafson and Kessel [5], Hathaway et al. [6], Honda and Ichihashi [7], Husseinzadeh Kashan et al. [9], Miyamoto et al. [15], Pedrycz [17], Pedrycz and Bargiela [18], Wu and Yang [22], Yang et al. [25], and Yu and Yang [26]). Regarding the generalization of FCM clustering, Yu and Yang [26] proposed a generalized FCM (GFCM) model to unify numerous variations of FCM. However, initializations affect FCM clustering and its generalizations. In this paper, we evaluated a bias-correction approach by using an updating equation to adjust the effects of initial values and then propose the bias-correction fuzzy clustering methods.
The rest of this paper is organized as follows. Section 2 presents a brief review of the FCM and GFCM algorithms. Section 3 presents the procedures involved in deriving the bias-correction fuzzy clustering algorithms. In these procedures, a bias-correction term is first assessed using an updating equation. The bias-correction FCM (BFCM), inter-cluster separation (ICS), and Gustafson and Kessel (GK) algorithms are then proposed. The bias-correction term is used as the total information for fuzzy c-partitions so that the proposed BFCM, GK, and ICS algorithms can be used to adjust gradually the effects of poor initializations. Section 4 presents comparisons between different clustering algorithms. In the comparisons, the number of optimal clustering results, error rates and root mean squared errors (RMSEs) are used as performance evaluation criteria. Numerical and real data sets are used to demonstrate the effectiveness and usefulness of the proposed bias-correction algorithms. Finally, conclusions and discussion are stated in Section 5.

2. Fuzzy clustering algorithms

Let \(X = \{x_1, \ldots, x_n\}\) be a set of \(n\) data points in an \(s\)-dimensional real Euclidean space. Let \(c\) be a positive integer greater than one. The FCM objective function [2,23] is expressed as follows:

\[
J_m(\mu, a) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m ||x_k - a_i||^2
\]

(1)

where \(m > 1\) is the weighting exponent, \(a = \{a_1, \ldots, a_c\}\) is the set of cluster centers, and the membership \(\mu_{ik}\) represents the degree to which the data point \(x_k\) belongs to a cluster \(i\) with

\[
\mu = [\mu_{ik}]_{i<n} \in \mathbb{R}^{n \\ c} = \left\{ \mu = [\mu_{ik}]_{i<n} \mid \sum_{i=1}^{c} \mu_{ik} = 1, \mu_{ik} \geq 0, 0 < \sum_{k=1}^{n} \mu_{ik} < n \right\}
\]

The FCM algorithm is developed with the objective of obtaining a partition matrix \(\mu = [\mu_{ik}]_{i<n}\) and a set \(a = \{a_1, \ldots, a_c\}\) of cluster centers to minimize the objective function \(J_m(\mu, a)\). By Lagrange multiplier, the necessary conditions for the minimum of \(J_m(\mu, a)\) are the following updating equations:

\[
a_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m x_k}{\sum_{k=1}^{n} \mu_{ik}^m}
\]

(2)

\[
\mu_{ik} = \frac{||x_k - a_i||^2}{\sum_{j=1}^{c} ||x_k - a_j||^2}
\]

(3)

According to Eqs. (2) and (3), the FCM algorithm can be described as follows:

**FCM algorithm**

Step 1: Fix \(2 \leq c \leq n\) and fix any \(\varepsilon > 0\)

Give an initial \(a^{(0)}\) and let \(t = 0\).

Step 2: Compute the membership \(\mu^{(t+1)}\) with \(a^{(t)}\) using Eq. (3).

Step 3: Update the cluster center \(a^{(t+1)}\) with \(\mu^{(t+1)}\) using Eq. (2).

Step 4: Compare \(a^{(t+1)}\) to \(a^{(t)}\) in a convenient matrix norm \(\|\cdot\|\).

IF \(\|a^{(t+1)} - a^{(t)}\| < \varepsilon\), STOP

ELSE \(t = t + 1\) and return to step 2.

The FCM algorithm is the most commonly used clustering algorithm. Numerous generalizations of the FCM algorithm exist. Yu and Yang [26] proposed a unified model, called the generalized FCM (GFCM). The GFCM objective function is expressed as follows:

\[
f_m^p(\mu, a) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \left[ \mu_{ik}^m h_i(d(x_k, a_i)) - \frac{\gamma}{c} \sum_{j=1}^{c} h_0(d(a_i, a_j)) \right]
\]

(4)

where \(\sum_{i=1}^{c} \mu_{ik} = f_k \geq 0; \gamma > 0\) are constant weights; \(h_i(x), i = 0, 1, \ldots, c\) are continuous functions of \(x \in [0, +\infty)\) satisfying its derivative \(h_i'(x) > 0\) for all \(x \in [0, +\infty)\), and \(d(x_k, a_i)\) is the distance between the data point \(x_k\) and the cluster center \(a_i\). The GFCM framework enables modeling numerous FCM variants. By Lagrange multiplier, the necessary conditions for a minimum of \(f_m^p(\mu, a)\) are obtained as follows:
\begin{align}
  a_i &= \frac{\sum_{k=1}^{n} \mu_{ik} h_i(d(x_k, a_i)) x_k - \frac{2}{\gamma} \sum_{j=1}^{c} h_0(d(a_i, a_j)) a_j}{\sum_{k=1}^{n} \mu_{ik} h_i(d(x_k, a_i)) - \frac{2}{\gamma} \sum_{j=1}^{c} h_0(d(a_i, a_j))} \\
  \mu_{ik} &= f_k \left( \frac{h_i(d(x_k, a_i))}{\sum_{j=1}^{c} h_i(d(x_k, a_j))} \right)^{\frac{1}{\gamma}}
\end{align}

The iterations with updating Eqs. (5) and (6) are called the GFCM algorithm. Although the FCM and GFCM algorithms return excellent clustering results when optimal initial values are provided, these fuzzy clustering algorithms are always affected by initial values. An example illustrating this phenomenon is presented in the next section.

3. Bias-correction fuzzy clustering algorithms

In general, the FCM and GFCM algorithms are affected by initializations; that is, the FCM and GFCM algorithms may return poor clustering results when poor initializations are used. Therefore, to overcome this drawback of the FCM algorithm and its generalizations, such as the GFCM algorithm, we propose a bias-correction term for reducing the effects of poor initializations. First, consider a probability mass \( p_i \) for the cluster center \( a_i, \ i = 1, \ldots, c \) with \( \sum_{i=1}^{c} p_i = 1 \); the probability mass \( p_i \) can be used to represent the proportion of the cluster center \( a_i \) to the \( c \) clusters. Theoretically, the term \( -\ln(p_i) \) can represent the information on the occurrence of the cluster center \( a_i \). The total information based on fuzzy c-partitions \( \mu_{ik} \) can thus be expressed as \( -\sum_{i=1}^{c} \sum_{j=1}^{c} \mu_{ik} \ln(p_i) \), which can be denoted as entropy. An optimal \( p_i \) can be determined by minimizing the entropy to obtain the most information for \( p_i \). In general, \(-w\sum_{i=1}^{c} \sum_{j=1}^{c} \mu_{ik} \ln(p_i)\) is used as a bias correction term to the GFCM objective function shown in Eq. (4) as follows:

\[ J_m(\mu, a, p) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} h_i(d(x_k, a_i)) - \frac{1}{\gamma} \sum_{j=1}^{c} h_0(d(a_i, a_j)) \]

\[ -w\sum_{k=1}^{n} \sum_{j=1}^{c} \mu_{ik} \ln(p_i) \]

By Lagrange multiplier, we can get the necessary conditions for minimum of \( J_m(\mu, a, p) \) as follows:

\[ a_i = \frac{\sum_{k=1}^{n} \mu_{ik} h_i(d(x_k, a_i)) x_k - \frac{2}{\gamma} \sum_{j=1}^{c} h_0(d(a_i, a_j)) a_j}{\sum_{k=1}^{n} \mu_{ik} h_i(d(x_k, a_i)) - \frac{2}{\gamma} \sum_{j=1}^{c} h_0(d(a_i, a_j))} \]

\[ \mu_{ik} = f_k \left( \frac{h_i(d(x_k, a_i))}{\sum_{j=1}^{c} h_i(d(x_k, a_j))} \right)^{\frac{1}{\gamma}} \]

\[ p_j = \frac{\sum_{i=1}^{n} \mu_{ik}}{\sum_{k=1}^{n} \sum_{j=1}^{c} \mu_{ik}} \]

For Eq. (9), we determine that, if \( w \to \infty \), then \( \mu_{ik} = f_k \left( \frac{-\ln(p_i)}{\sum_{j=1}^{c} (-\ln(p_j))} \right)^{\frac{1}{\gamma}} \). If \( w \to 0 \), then \( \mu_{ik} = f_k \left( \frac{h_i(d(x_k, a_i))}{\sum_{j=1}^{c} h_i(d(x_k, a_j))} \right)^{\frac{1}{\gamma}} \) as shown in Eq. (6) of the GFCM algorithm. Therefore, an updating equation may be used for the parameter \( w \) with

\[ w^{(t)} = (0.99)^t \]

where \( t \) is the number of iterations. Thus, the iterated algorithm with updating Eqs. (8)-(10) and decreasing learning rate of the updating Eq. (11) are called the bias-correction GFCM algorithm. Next, the three types of bias-correction GFCM algorithms, which are the BFCM, bias-correction GK (BGK), and bias-correction ICS (BICS) algorithms, are assessed.

3.1. Bias-correction FCM algorithm

The bias-correction FCM (BFCM) objective function is expressed as follows:

\[ J_m(\mu, a, p) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik} ||x_k - a_i||^2 - w \sum_{k=1}^{n} \sum_{j=1}^{c} \mu_{ik} \ln(p_i) \]

subject to \( \sum_{i=1}^{c} h_i = 1 \) and \( \sum_{j=1}^{c} p_i = 1 \). Thus, the BFCM algorithm is iterated under the necessary conditions by using the following updating equations:

\[ a_i = \frac{\sum_{k=1}^{n} \mu_{ik} x_k}{\sum_{k=1}^{n} \mu_{ik}} \]

\[ \mu_{ik} = \frac{(||x_k - a_i||^2 - w \ln(p_i))^{\frac{1}{\gamma}}}{\sum_{j=1}^{c} (||x_k - a_j||^2 - w \ln(p_j))^{\frac{1}{\gamma}}} \]

\[ p_j = \frac{\sum_{i=1}^{n} \mu_{ik}}{\sum_{k=1}^{n} \sum_{j=1}^{c} \mu_{ik}} \]
Similarly, we also consider the same decreasing learning rate with the updating Eq. (11) for the parameter \( w \). The BFCM algorithm can be summarized as follows:

**BFCM algorithm**

Step 1: Fix \( 2 \leq c \leq n \) and fix any \( \varepsilon > 0 \). Give an initial \( a^{(0)}, p^{(0)} = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right) \) and let \( t = 0, w^{(0)} = 1 \).

Step 2: Learn the parameter \( w^{(t)} \) using Eq. (11).

Step 3: Compute the membership \( \mu^{(t+1)} \) with \( a^{(t)} \) and \( p^{(t)} \) using Eq. (14).

Step 4: Compute the probability weight \( p_{ij}^{(t+1)} \) using Eq. (15);

Step 5: Update the cluster center \( a^{(t+1)} \) with \( \mu^{(t+1)} \) using Eq. (13).

Step 6: Compare \( a^{(t+1)} \) to \( a^{(t)} \) in a convenient matrix norm \( \| \cdot \| \).

IF \( \| a^{(t+1)} - a^{(t)} \| < \varepsilon \), STOP.

ELSE \( t = t + 1 \) and return to step 2.

3.2. Bias-correction GK algorithm

Obviously, using the Euclidean distance as a distance measure can lead to optimal results only when a data set containing spherical clusters is used. The FCM algorithm is not ideal for analyzing a data set containing clusters with different shapes. To overcome this drawback, Gustafson and Kessel [5] assessed the effects of different cluster shapes by replacing the Euclidean (squared) distance \( d(x_i, x_j) = \| x_i - x_j \|^2 \) in the FCM algorithm with the Mahalanobis distance \( d(x_i, x_j) = \| x_i - x_j \|^2_{A_j} = (x_i - x_j)^T A_j (x_i - x_j) \) and proposed the GK algorithm. The GK objective function is expressed as follows:

\[
J_m^GK(\mu, A) = \sum_{j=1}^{n} \sum_{i=1}^{c} \mu_{ij}^m \| x_i - a_j \|^2_{A_j}
\]

(16)

with \( \mu \in M_{mn}, a = (a_1, \ldots, a_c) \in \mathbb{R}^d \) and \( A = \{ A_1, \ldots, A_c \} \) which \( A_i \) is positive definite with \( \det(A_i) = \rho_i \). The necessary conditions for minimizing \( J_m^GK(\mu, A) \) are the following updating equations:

\[
a_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m x_k}{\sum_{k=1}^{n} \mu_{ik}^m}
\]

(17)

\[
\mu_{ik} = \frac{\| x_i - a_k \|_{A_k}^{-2/(m-1)}}{\sum_{j=1}^{c} \| x_i - a_j \|_{A_j}^{-2/(m-1)}}
\]

(18)

with \( A_i = (\rho_i \det(S_i))^{1/2} S_i^{-1}, S_i = \sum_{k=1}^{n} \mu_{ik}^m (x_k - a_i)(x_k - a_i)^T / \sum_{k=1}^{n} \mu_{ik}^m \) \( i = 1, \ldots, c \). Thus, the GK algorithm can be created according to the updating Eqs. (17) and (18).

By adding the bias-correction term \( -w \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \ln(p_i) \), the BGK objective function can be expressed as follows:

\[
J_m(\mu, A, p) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \| x_k - a_i \|^2_{A_i} - w \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \ln(p_i)
\]

(19)

subject to \( \sum_{k=1}^{c} u_{ik} = 1 \) and \( \sum_{j=1}^{n} p_j = 1 \). Thus, the BGK algorithm is iterated under the necessary conditions by using the following updating equations:

\[
a_i = \frac{\sum_{j=1}^{n} \mu_{ij}^m x_j}{\sum_{j=1}^{n} \mu_{ij}^m}
\]

(20)

\[
p_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m}{\sum_{j=1}^{n} \sum_{k=1}^{n} \mu_{jk}^m}
\]

(21)

\[
\mu_{ik} = \left( \frac{\| x_i - a_k \|_{A_k}^2 - w \ln(p_i)}{\sum_{j=1}^{c} (\| x_i - a_j \|_{A_j}^2 - w \ln(p_j))} \right)^{-1/(m-1)}
\]

(22)

with \( A_i = (\rho_i \det(S_i))^{1/2} S_i^{-1}, S_i = \sum_{k=1}^{n} \mu_{ik}^m (x_k - a_i)(x_k - a_i)^T / \sum_{k=1}^{n} \mu_{ik}^m \) \( i = 1, \ldots, c \). Similarly, in this study, we used the updating Eq. (11) with the \( w \) parameter. Thus, the BGK algorithm can be summarized as follows:
### BICS algorithm

Step 1: Fix $2 \leq c \leq n$ and fix any $\epsilon > 0$

Given an initial $a^{(0)}, p^{(0)} = \left(\frac{1}{c}, \frac{1}{c}, \ldots, \frac{1}{c}\right)$ and let $t = 0, w(0) = 1$.

Step 2: Learn the parameter $w^{(t)}$ using Eq. (11).

Step 3: Compute the membership $\mu^{(t+1)}$ with $a^{(t)}$ and $p^{(t)}$ using Eq. (29).

Step 4: Compute the probability weight $p_i^{(t+1)}$ using Eq. (21);

Step 5: Update the cluster center $a_i^{(t+1)}$ with $\mu^{(t+1)}$ using Eq. (20).

Step 6: Compare $a_i^{(t+1)}$ to $a_i^{(t)}$ in a convenient matrix norm $\|\|$. IF $\|a^{(t+1)} - a^{(t)}\| < \epsilon$, STOP.

ELSE $t = t + 1$ and return to step 2.

### 3.3. Bias-correction ICS algorithm

For the GFCM objective function $J^h_m(\mu, a)$, if we have $h_i(d(x_k, a_j)) = ||x_k - a_j||^2$ and $h_0(d(a_j, a_j)) = ||a_i - a_j||^2$, then we can obtain the objective function of inter-cluster separation (ICS) fuzzy clustering proposed by Özdemir and Akarun [16] as follows:

$$J^ICS_m(\mu, a) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \left( \mu_{ik}^m ||x_k - a_i||^2 - \frac{c}{c} \sum_{j=1}^{c} ||a_i - a_j||^2 \right), \gamma \geq 0$$

(23)

The necessary conditions for minimizing $J^ICS_m(\mu, a)$ are the following updating equations:

$$a_i = \frac{\frac{1}{n} \sum_{k=1}^{n} \mu_{ik}^m ||x_k - a_i||^2 - \frac{c}{c} \sum_{j=1}^{c} a_j}{\frac{1}{n} \sum_{k=1}^{n} \mu_{ik}^m ||x_k - a_i||^2 - 2\gamma}$$

(24)

$$\mu_{ik} = \frac{||x_k - a_i||^2}{\sum_{j=1}^{c} ||x_k - a_j||^2}$$

(25)

By adding the bias-correction term $(-w\sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \ln(p_i))$, the bias-correction ICS (BICS) objective function can be expressed as follows:

$$J^{BICS}_m(\mu, a, p) = \frac{1}{n} \sum_{k=1}^{n} \sum_{i=1}^{c} \left( \mu_{ik}^m ||x_k - a_i||^2 - \frac{c}{c} \sum_{j=1}^{c} ||a_i - a_j||^2 \right) - w \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m \ln(p_i), \gamma \geq 0$$

(26)

The necessary conditions for minimizing $J^{BICS}_m(\mu, a, p)$ are the following updating equations:

$$a_i = \frac{\frac{1}{n} \sum_{k=1}^{n} \mu_{ik}^m ||x_k - a_i||^2 - \frac{c}{c} \sum_{j=1}^{c} a_j}{\frac{1}{n} \sum_{k=1}^{n} \mu_{ik}^m ||x_k - a_i||^2 - 2\gamma}$$

(27)

$$\mu_{ik} = \left( \frac{||x_k - a_i||^2 - w \ln(p_i)}{\sum_{j=1}^{c} ||x_k - a_j||^2 - w \ln(p_j)} \right)$$

(28)

$$p_i = \frac{\sum_{j=1}^{c} \mu_{ij}^m}{\sum_{j=1}^{c} \mu_{jk}^m}$$

(29)

Thus, the BICS algorithm can be summarized as follows:

### BICS algorithm

Step 1: Fix $2 \leq c \leq n$ and fix any $\epsilon > 0$

Given an initial $a^{(0)}, p^{(0)} = \left(\frac{1}{c}, \frac{1}{c}, \ldots, \frac{1}{c}\right)$ and let $t = 0, w(0) = 1$.

Step 2: Learn the parameter $w^{(t)}$ using Eq. (11).

Step 3: Compute the membership $\mu^{(t+1)}$ with $a^{(t)}$ and $p^{(t)}$ using Eq. (29);

Step 4: Compute the probability weight $p_i^{(t+1)}$ using Eq. (28);

Step 5: Update the cluster center $a_i^{(t+1)}$ with $\mu^{(t+1)}$ using Eq. (27).

Step 6: Compare $a_i^{(t+1)}$ to $a_i^{(t)}$ in a convenient matrix norm $\|\|$. IF $\|a^{(t+1)} - a^{(t)}\| < \epsilon$, STOP.

ELSE $t = t + 1$ and return to step 2.
The main differences between the fuzzy clustering algorithms, such as FCM, ICS and GK, and the proposed bias-correction fuzzy clustering algorithms, such as BFCM, BICS and BGK, are that in the proposed bias-correction fuzzy clustering algorithms, the probability mass $p_i$ for the cluster center $a_i$ is calculated using the equation $p_i = \frac{\sum_{k=1}^{n} \mu_{ik}^m}{\sum_{k=1}^{n} \sum_{i=1}^{m} \mu_{ik}^m}$ and the fuzzy c-partition $\mu_{ik}$ is updated by replacing the distance between the data point $x_k$ and the cluster center $a_i$, such as $||x_k - a_i||^2$, with $(\text{distance} - w \ln(p_i))$, such as $||x_k - a_i||^2 - w \ln(p_i)$. Furthermore, in the proposed bias-correction fuzzy clustering algorithms, an updating equation is used for the $w$ parameter with $w^{t+1} = (0.99)^t$ to ensure that the bias-correction decreases after more iteration steps are conducted. In this sense, the proposed bias-correction fuzzy clustering algorithms can gradually adjust the bias caused by the effects of poor initializations. We subsequently conducted experiments to demonstrate the robustness of the proposed bias-correction fuzzy clustering algorithms to initializations.

Fig. 1. Ten-cluster data set where “*” denotes the true cluster centers.

Fig. 2. Trajectories of initializations for FCM, where the numbers 1–10 denote 10 initializations and “*” denotes the true cluster centers.

Fig. 3. Trajectories of initializations for BFCM, where the numbers 1–10 denote 10 initializations and “*” denotes the true cluster centers.
Fig. 4. (a) Nine-cluster data set with 9 initial cluster centers; (b)–(j) cluster centers after 1, 5, 10, 20, 30, 40, 50, 60 and 70 iterations, respectively, from FCM; (k) convergent cluster centers after 74 iterations from FCM; and (l) final clustering results from FCM.
3.4. Analysis on influence of initializations

We examined the correction behavior of the bias-correction term \(-w^t \sum_{i=1}^{n} \sum_{j=1}^{c} \mu_{ij}^t \ln(p_j)\) for initializations. We observed the bias-correction term in mathematical form and derived a term with a positive value to ensure that the BFCM objective function \(J_m(\mu, a, p)\) is greater than the FCM objective function \(J_m(\mu, a)\). However, the bias-correction term presents the total information on the occurrence of the cluster centers with weighted fuzzy c-partitions. This is analogous to heating the cost function so that it becomes more flexible for adjusting the poor initializations and then cools down the system by a decreasing learning rate of \(w^t = (0.99)^t\). To analyze the influence of initializations, we present some examples to demonstrate the behavior. Because the BFCM, BGK, and BICS algorithms all involve adding the same bias-correction term, we present initialization analysis only for the BFCM algorithm. We first present the correction behavior for the bias-correction term by using the moving routes of cluster centers during iterations in Example 1. We then present more examples that involve analyses of the correction behavior of the bias-correction term by using mean squared errors (MSEs).

**Example 1.** We used a simulated data set in this example. As illustrated in Fig. 1, the data set comprised 10 clusters and each cluster comprised 100 data points; “•” denotes the true cluster centers. We choose the 10 initial cluster centers \(a_1^0 = (0, 7.5), a_2^0 = (2, 7.5), a_3^0 = (4, 7.5), a_4^0 = (6, 7.5), a_5^0 = (8, 7.5), a_6^0 = (0, 0), a_7^0 = (3.5, 0), a_8^0 = (7, 0), a_9^0 = (0, 1.5), a_{10}^0 = (7, 1.5)\) for both FCM and BFCM algorithms, where the numbers 1–10 denotes 10 initializations, as shown in Figs. 2 and 3. Fig. 2 depicts the initialization trajectories for the FCM algorithm, indicating that the two initializations at \(a_1^0 = (0, 7.5)\) and \(a_2^0 = (2, 7.5)\) approach the same cluster center; therefore, no initialization approached the true cluster center \((3.5, 0.5)\). Fig. 3 shows the initialization trajectories for the BFCM algorithm, indicating that some routes are affected by the bias correction; therefore, all initializations finally approached the true cluster center. These trajectories demonstrate the effects of initializations on the bias-correction term. In summary, the BFCM algorithm is actually more robust against initializations than the FCM is.

**Example 2.** We used a simulated data set in this example as well. The data set comprised nine clusters and each cluster has 100 data points, as shown in Figs. 4(a) and 6(a). We choose the 9 initial cluster centers \(a_1^0 = (3, 4, 1.6), a_2^0 = (3.3, 0.6), a_3^0 = (2.6, 2.3), a_4^0 = (0.4, 4.3), a_5^0 = (4.6, 2.0), a_6^0 = (3.4, 0.5), a_7^0 = (5.1, 2.1), a_8^0 = (3.4, 0.9)\) and \(a_9^0 = (0.1, 2.1)\) for both FCM and BFCM algorithms, as shown in Figs. 4(a) and 6(a). Fig. 4 illustrates the centers of these clusters after 1, 10, 30, 60, and 74 iterations in the FCM algorithm; the final clustering results are also shown in this figure. Fig. 5 shows the plot of the MSEs of the FCM algorithm for different iterations. Furthermore, Fig. 6 illustrates the centers of these clusters after 1, 10, 20, 30, 40, 50, 60, 100, 500 and 507 iterations in the BFCM algorithm as well as the final clustering results. Fig. 7 depicts the plot of the MSEs of the BFCM algorithm for different iterations. The plots of the MSEs of both the FCM and BFCM algorithms for the first 70 iterations were incorporated (Fig. 8) for comparison. As shown in Figs. 4, 5, and 8, the FCM algorithm adjusts initializations at an extremely fast rate, thus preventing it from targeting actual cluster centers. The BFCM algorithm gradually adjusts initializations and then targets the true cluster centers after the iteration \(t = 60\) (Figs. 6–8). This behavior indicates the effects of adjusting the initialization on the bias-correction term. In general, the BFCM algorithm is actually more robust against initializations than the FCM algorithm is.

This example is based on Example 2. However, we use nine additional initial cluster centers with \(a_1^0 = (1.1, 4.7), a_2^0 = (4.2, 4.1), a_3^0 = (1.1, 2.8), a_4^0 = (4.8, 2.6), a_5^0 = (1.6, 4.6), a_6^0 = (0.8, 1.2), a_7^0 = (3.1, 3.7), a_8^0 = (2.3, 1.7)\) and \(a_9^0 = (1.6, 2.7)\) for both FCM and BFCM algorithms. We determined that both the FCM and BFCM algorithms returned...
Fig. 6. (a) Nine-cluster data set with 9 initial cluster centers; (b)–(j) cluster centers after 1, 10, 20, 30, 40, 50, 60, 100 and 500 iterations, respectively, from BFCM; (k) convergent cluster center after 507 iterations from BFCM; and (l) final clustering results from BFCM.
Fig. 7. Plot of MSEs from BFCM.

Fig. 8. Comparisons between MSEs from FCM and BFCM for the first 70 iterations.

Fig. 9. MSE results from FCM and BFCM with various initializations.
optimal final clustering results with extremely low MSEs (Fig. 9, fifth point). The final clustering results of the FCM algorithm may be changeable, even if the initialization has a small change. However, the BFCM algorithm still has stable final clustering results with an extremely low MSE. These procedures were conducted by assigning various values to the second component of $a_1^{(0)}$, as shown in Table 1. For example, $a_1^{(0)} = (1.1, 0.7)$ is the first point and $a_1^{(0)} = (1.1, 4.7)$ is the fifth point with their respective MSEs shown in Table 1 and Fig. 9. As shown in Table 1 and Fig. 9, the FCM algorithm is sensitive to initializations and its final clustering results can be changed. However, the BFCM algorithm is not sensitive to initializations and its final clustering results are stable.

In this example, we studied different learning behaviors for the parameter $w$. We executed the BFCM algorithm for the data set employed in Example 2. We assessed the learning approach by using the updating equation $w = (0.9)^t$, and the results indicated that initializations are adjusted at an extremely fast rate, thus preventing them from targeting optimal cluster centers, as shown in Fig. 10(a). Similarly, the clustering results (Fig. 10(e) and (f)) obtained using the updating equations $w = (0.99)^t$ and $w = (0.999999)^t$, respectively, are not optimal; this is because initializations are adjusted at an extremely slow rate, thus preventing them from approaching optimal clustering centers. Regarding the clustering results (Fig. 10(b)–(d)) obtained using the updating equations $w = (0.99)^t$, $w = (0.999)^t$ and $w = (0.9999)^t$, initializations can be adjusted and targeted to optimal clustering centers. However, the MSE of the updating equation $w = (0.99)^t$ is lower than that of $w = (0.999)^t$ and $w = (0.999999)^t$, as shown in Table 2 and Fig. 11. In general, we recommend using the updating equation $w = (0.99)^t$ as a decreasing learning approach for the parameter $w$ in the bias-correction term.

### Table 1
MSEs of FCM and BFCM with different initializations in $a_1^{(0)}$.

<table>
<thead>
<tr>
<th>$a_1^{(0)}$</th>
<th>(1.1, 0.7)</th>
<th>(1.1, 1.7)</th>
<th>(1.1, 2.7)</th>
<th>(1.1, 3.7)</th>
<th>(1.1, 4.7)</th>
<th>(1.1, 5.7)</th>
<th>(1.1, 6.7)</th>
<th>(1.1, 7.7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>0.3417</td>
<td>0.0007</td>
<td>0.5183</td>
<td>0.0007</td>
<td>0.6210</td>
<td>0.6207</td>
<td>0.8709</td>
<td></td>
</tr>
<tr>
<td>BFCM</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
<td>0.0007</td>
</tr>
</tbody>
</table>
3.5. Convergence properties of the BFCM clustering algorithm

We assessed a convergence theorem for the proposed BFCM clustering algorithm; that is, we can guarantee that any BFCM convergent subsequence will tend to optimal solutions. Zangwill’s convergence theorem [27] must be applied first. Zangwill [27] originally defined a point-to-set map as $T : V \rightarrow P(V)$, where $P(V)$ represents the power set of $V$; a closed point-to-set map must be defined. However, the algorithms of interest here is a point-to-point map and the “closed” property is exactly “continuity” for the case of point-to-point map. Thus, the Zangwill’s convergence theorem [27] is given as follows:

**Zangwill’s convergence theorem** [27]. Let the point-to-point map $T : V \rightarrow V$ generate a sequence $\{z_k\}_{k=0}^\infty$ by $z_{k+1} = T(z_k)$. Let a solution set $\Omega \subset V$ be given, and suppose that:

1. There is a continuous function $Z : V \rightarrow R$ such that
   - if $z \notin \Omega$, then $Z(T(z)) < Z(z)$, and
   - if $z \in \Omega$, then $Z(T(z)) \leq Z(z)$.
2. The map $T$ is continuous on $V \setminus \Omega$.
3. All point $z_k$ are contained in a compact set $S \subset V$.

Then the limit of any convergent subsequence shall be in the solution set $\Omega$, and $Z(z_k)$ will monotonically converge to $Z(z)$ for some $z \in \Omega$.

We set that $M_{Fcm} = \{ \mu = [\mu_{ik}]_{i \in \mathbb{N}} \mid \sum_{i=1}^n \mu_{ik} = 1, \mu_{ik} \geq 0, 0 < \sum_{k=1}^n \mu_{ik} < n \}$, $M_p = \{ p = [p_i]_{i \in \mathbb{N}} \mid \sum_{i=1}^c p_i = 1, p_i \geq 0 \}$ and $a = (a_1^T, \ldots, a_c^T)^T$. Let

$$\Omega_\epsilon = \left\{ (\mu^*, a^*, p^*) \left\| \begin{array}{l}
\forall \mu \in M_{Fcm}, \forall \mu^* \neq \mu, \sum_{i=1}^c \mu^*_{ik} < \sum_{i=1}^c \mu_{ik} \\
\forall a \neq a^*, \sum_{i=1}^c a^*_{ik} < \sum_{i=1}^c a_{ik} \\
\forall p \neq p^*, \sum_{i=1}^c p^*_{ik} < \sum_{i=1}^c p_{ik}
\end{array} \right. \right\}$$

where $\mu^* = [\mu_{ik}]_{i \in \mathbb{N}}$ with $\mu_{ik}^* = \frac{1}{\sum_{p_{ik}^*}^n (||x_i - a_{ik}||^2 - \ln(p_{ik}^*))^{1/2}}$, $a^* = (a_1^T, \ldots, a_c^T)^T$ with $a_i^T = \frac{\sum_{k=1}^c \mu_{ik}^* a_{ik}}{\sum_{k=1}^c \mu_{ik}^*}$ and $p^* = [p_i]_{i \in \mathbb{N}}$ with $p_i = \frac{1}{\sum_{k=1}^c (||x_i - a_{ik}||^2 - \ln(p_{ik}^*))^{1/2}}$. Let $E : (R^T)^c \times M_p \rightarrow M_{Fcm}$ with $E(a, p) = \mu = [\mu_{ik}]_{i \in \mathbb{N}}$, where $\mu_{ik}$ is calculated via $\mu_{ik} = \frac{1}{\sum_{k=1}^c (||x_i - a_{ik}||^2 - \ln(p_{ik}^*))^{1/2}}$. Let

**Table 2**

MSEs of BFCM and iteration number in various updating equations of $w$.

<table>
<thead>
<tr>
<th>$w$</th>
<th>(0.9)$^1$</th>
<th>(0.99)$^1$</th>
<th>(0.999)$^1$</th>
<th>(0.9999)$^1$</th>
<th>(0.99999)$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.5642</td>
<td>0.0007</td>
<td>0.0011</td>
<td>0.0416</td>
<td>0.2245</td>
</tr>
<tr>
<td>iteration</td>
<td>78</td>
<td>329</td>
<td>1933</td>
<td>5268</td>
<td>224</td>
</tr>
</tbody>
</table>

**Fig. 11.** MSEs of BFCM in various updating equations of $w$. 
\[ F : M_{\text{fcm}} \rightarrow (R^r)^\ast, F(\mu) = a = (a_1^T, \ldots, a_n^T)^T, \] where \( a_i \) is calculated via \( a_i = \sum_{r=1}^n \frac{\partial f_k}{\partial x_r} \). Let \( G : M_{\text{fcm}} \rightarrow M_p, G(\mu) = p = |p|_{k \times 1} \) where \( p_i \) is calculated by \( p_i = \sum_{r=1}^n \frac{\partial p_k}{\partial \mu_r} \). We next define the BFCM operator as follows.

**Definition 1.** The BFCM operator \( T_m : M_{\text{fcm}} \times (R^r)^\ast \times M_p \rightarrow M_{\text{fcm}} \times (R^r)^\ast \times M_p \) is defined by \( T_m = A_2 \circ A_1 \) where \( A_1 : M_{\text{fcm}} \times (R^r)^\ast \times M_p \rightarrow M_{\text{fcm}} \) with \( A_1(\mu, a, p) = E(a, p) \) and \( A_2 : M_{\text{fcm}} \rightarrow (R^r)^\ast \times M_p \) with \( A_2(\mu) = (\mu, F(\mu), G(\mu)) \).

Thus, we have
\[
T_m(\mu, a, p) = (A_2 \circ A_1)(\mu, a, p) = A_2(A_1(\mu, a, p)) = A_2(E(a, p)) = (E(a, p), F(E(a, p)), G(E(a, p))) = (\mu', a', p')
\]
where \( \mu' = E(a, p), a' = F(E(a, p)) = F(\mu') \text{ and } p' = G(E(a, p)) = G(\mu') \).

In general, the necessary and sufficient condition for a strict minimizer of an objective function is to analyze the Jacobian matrix and Hessian matrix. However, if some constraints are considered, Lagrange’s multipliers in addition to a bordered Hessian matrix must be assessed as follows.

**Theorem 1** (Lagrange’s theorem; see Werner and Sotskov [21], pp. 425–426). Let functions \( f : D_f \rightarrow R, D_f \subseteq R^n \), and \( g_i : D_g_i \rightarrow R, D_g_i \subseteq R^n, i = 1, \ldots, t < n \), be continuously partially differentiable and let \( x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \in D_f \) be a local extreme point of the function \( f \) subject to the constraints \( g_i(x_1, x_2, \ldots, x_n) = 0, i = 1, 2, \ldots, t \). Let \( L(x, \lambda) = f(x_1, x_2, \ldots, x_n) + \sum_{i=1}^t \lambda_i g_i(x_1, x_2, \ldots, x_n) \) and
\[
J = \begin{vmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{vmatrix} \neq 0
\]
at the point \( x^0 \). Then, we have that the gradient of \( L(x, \lambda) \) at the point \( (x^0, \lambda^0) \) is 0, i.e., \( \nabla L(x^0, \lambda^0) = 0 \).

**Theorem 2** (local sufficient conditions; see Werner and Sotskov [21], pp. 426–427). Let functions \( f : D_f \rightarrow R, D_f \subseteq R^r, \) and \( g_i : D_g_i \rightarrow R, D_g_i \subseteq R^n, i = 1, \ldots, t < n \), be twice continuously partially differentiable and let \( (x^0, \lambda^0) \) with \( x^0 \in D_f \) be a solution of the system \( \nabla L(x^0, \lambda^0) = 0 \). Let
\[
H_f(x, \lambda) = \begin{bmatrix} 0 & \cdots & 0 & \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} \\ \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} & \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} & \frac{\partial^2 f}{\partial x_1 \partial \lambda_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial \lambda_n} \end{bmatrix}
\]
be the bordered Hessian and consider its leading principle minors \( |\overline{H}_i(x^0, \lambda^0)| \) of the order \( r = 2t + 1, 2t + 2, \ldots, n + t \) at point \( (x^0, \lambda^0) \). Therefore, the following expressions can be derived:

(1) If all leading principle minors, \( |\overline{H}_i(x^0, \lambda^0)|, 2t + 1 \leq r \leq n + t \), have the sign \((-1)^i\), then \( x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \) is a local minimum point of function \( f \) subject to the constraints \( g_i(x) = 0, i = 1, 2, \ldots, t \).

(2) If the signs of all leading principle minors \( |\overline{H}_i(x^0, \lambda^0)|, 2t + 1 \leq r \leq n + t \), are alternated, the sign of \( |\overline{H}_{n+i}(x^0, \lambda^0)| \) being that of \((-1)^i\), then \( x^0 = (x_1^0, x_2^0, \ldots, x_n^0) \) is a local maximum point of function \( f \) subject to the constraints \( g_i(x) = 0, i = 1, 2, \ldots, t \).

(3) If neither the conditions of (1) nor those of (2) are satisfied, then \( x^0 \) is not a local extreme point of function \( f \) subject to the constraints \( g_i(x) = 0, i = 1, 2, \ldots, t \). Here, the case in which one or several leading principal minors have a value of zero is not considered a violation of condition (1) or (2).

**Remark 1** (see Werner and Sotskov [21], p. 379). The leading principal minors of matrix \( A = (a_{ij}) \) of order \( n \times n \) are the determinants
\[ D_k = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}, \quad k = 1, 2, \ldots, n \]

i.e. \( D_k \) is obtained from \( |A| \) by crossing out the last \( n - k \) columns and rows.

**Lemma 1.** If \( \mu = \hat{\mu} \) and \( p = \hat{p} \) are fixed, then \( J_m(\mu, a, p) \) is minimized at \( a^* = (a_1^T, \ldots, a_c^T)^T \) if and only if
\[
a_i^* = \frac{\sum_{k=1}^n \mu_{ik}^{(m)}}{\sum_{k=1}^n \mu_{ik}^{(m)}}, \quad \forall i = 1, \ldots, c.
\]

**Proof.** Recall that \( J_m(\mu, a, p) = \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^{(m)} \|x_k - a_i\|^2 - w \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^{(m)} \ln(p_i) \). With the gradient of \( J_m \) w.r.t. \( a_i \), we have
\[
\frac{\partial J_m}{\partial a_i} = \sum_{k=1}^n \mu_{ik}^{(m)}(-2x_k - a_i)) = 0 \Rightarrow a_i^* = \frac{\sum_{k=1}^n \mu_{ik}^{(m)}x_k}{\sum_{k=1}^n \mu_{ik}^{(m)}}, \quad \forall i = 1, \ldots, c
\]

Thus, we had proved the "only if" condition. We next prove the "if" condition as follows. If \( \mu = \hat{\mu} \) and \( p = \hat{p} \) are fixed, then we have
\[
\frac{\partial J_m}{\partial a_i} = \sum_{k=1}^n \mu_{ik}^{(m)}(-2x_k - a_i)) \quad \text{and} \quad \frac{\partial^2 J_m}{\partial a_i^2} = 2 \times \delta_{ij} \times \bar{I} \times \sum_{k=1}^n \mu_{ik}^{(m)} - \delta_{ij},
\]
where \( \delta_{ij} \) is Kronecker index with \( \delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \). Thus, the Hessian matrix of \( J_m(\mu, a, p) \) w.r.t. \( a \) is \( 2 \times \text{diag}(I \times \sum_{k=1}^n \mu_{ik}^{(m)}), I \times \sum_{k=1}^n \mu_{ik}^{(m)} \times \ldots \times I \times \sum_{k=1}^n \mu_{ik}^{(m)} \) and obviously, the Hessian matrix is positive definite. That is, \( J_m(\mu, a, p) \) is minimized at \( a^* = (a_1^T, \ldots, a_c^T)^T \) with \( a_i^* = \frac{\sum_{k=1}^n \mu_{ik}^{(m)}x_k}{\sum_{k=1}^n \mu_{ik}^{(m)}}, \quad \forall i = 1, \ldots, c \). \( \square \)

**Lemma 2.** If \( \mu = \hat{\mu} \) and \( a = \hat{a} \) are fixed, then \( J_m(\mu, a, p) \) subject to \( \sum_{i=1}^c p_i = 1 \) is minimized at \( p^* = [p_i^*] \) if and only if
\[
p_i^* = \frac{\sum_{k=1}^n \mu_{ik}^{(m)}}{\sum_{k=1}^n \mu_{ik}^{(m)}}, \quad \forall i = 1, \ldots, c.
\]

**Proof.** Let the Lagrangian function be
\[
L_1 = \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^{(m)} \|x_k - a_i\|^2 - w \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^{(m)} \ln(p_i) + \eta \left( \sum_{i=1}^c p_i - 1 \right)
\]
where \( \eta \) is a Lagrangian multiplier. With the gradient of \( L_1 \) w.r.t. \( p_i \) and \( \eta \), we have
\[
\begin{cases}
\frac{\partial L_1}{\partial p_i} = -w \sum_{k=1}^n \mu_{ik}^{(m)} \frac{1}{p_i} + \eta = 0 \\
\frac{\partial L_1}{\partial \eta} = \sum_{i=1}^c p_i - 1 = 0
\end{cases}
\]

Thus, we have \( p_i^* = \frac{\sum_{k=1}^n \mu_{ik}^{(m)}}{\sum_{i=1}^n \sum_{k=1}^n \mu_{ik}^{(m)}}, \quad \eta^* = \sum_{i=1}^n \mu_{ik}^{(m)} \) and the "only if" condition is proved. We next prove the "if" condition by Theorem 2 as follows. If \( \mu = \hat{\mu} \) and \( a = \hat{a} \) are fixed, then
\[
\frac{\partial L_1}{\partial p_i} = -w \sum_{k=1}^n \mu_{ik}^{(m)} \frac{1}{p_i} + \eta \frac{\partial \mu_{ik}^{(m)}}{\partial p_i} = \delta_{ij} \frac{\sum_{k=1}^n \mu_{ik}^{(m)}}{p_i^*} \quad \text{and} \quad \frac{\partial^2 L_1}{\partial p_i^2} = \frac{\partial^2 \mu_{ik}^{(m)}}{\partial p_i^2} = 1.
\]
Thus, the bordered Hessian matrix w.r.t. \( p \) and \( \eta \) is
\[
H_{L_1}(p, \eta) = \begin{bmatrix}
0 & 1 & 1 & \cdots & 1 \\
1 & 0 & \cdots & \cdots & 0 \\
1 & 0 & \cdots & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 1 & \cdots & \cdots & 0
\end{bmatrix}
\]
Note that we have only one constraint, i.e. \( t = 1 \), and so \((-1)^t = -1 < 0\). We next check all leading principle minors as follows:
\[
\begin{align*}
|H_3(p', \eta')| &= \begin{vmatrix}
0 & 1 & 1 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_1^2} & 0 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_2^2} & 0 & \cdots & 0 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_c^2} & 0 & \cdots & 0
\end{vmatrix}_{p-p', \eta-\eta'} = -\left( w\frac{\sum_{k=1}^{m} \mu_k^m}{p_1^2} + \frac{w\sum_{k=1}^{m} \mu_k^m}{p_2^2} \right)_{p-p', \eta-\eta'} < 0,

|H_4(p', \eta')| &= \begin{vmatrix}
0 & 1 & 1 & 1 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_1^2} & 0 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_2^2} & 0 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_c^2} & 0
\end{vmatrix}_{p-p', \eta-\eta'} = -\left( w^2 \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_1^2} \right) \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_1^2} \right) + \frac{w^2 \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_2^2} \right) \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_2^2} \right) + \frac{w^2 \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_c^2} \right) \left( \frac{\sum_{k=1}^{m} \mu_k^m}{p_c^2} \right)}{p_1^2 p_2^2 p_3^2} \right)_{p-p', \eta-\eta'} < 0
\end{align*}
\]

\[
|H_{c+1}(p', \eta')| = \begin{vmatrix}
0 & 1 & 1 & \cdots & 1 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_1^2} & 0 & \cdots & 0 \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_2^2} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & \frac{w\sum_{k=1}^{m} \mu_k^m}{p_c^2} & 0 & \cdots & 0
\end{vmatrix}_{p-p', \eta-\eta'} = -\left( \sum_{i=1}^{c} \prod_{j=1}^{c} \frac{w\sum_{k=1}^{m} \mu_k^m}{p_j^2} \right)_{j\neq i} < 0
\]

Thus, by Theorem 2, \( J_m(\mu, a, p) \) subject to \( \sum_{i=1}^{n} p_i = 1 \) is minimized at \( p^* = [p_i^*] \) with \( p_i^* = \frac{\sum_{j=1}^{c} \mu_k^m}{\sum_{j=1}^{c} \mu_k^m}, \forall i = 1, \ldots, n \). □

**Lemma 3.** If \( a = \hat{a} \) and \( p = \hat{p} \) are fixed, then \( J_m(\mu, \hat{a}, \hat{p}) \) subject to \( \sum_{i=1}^{c} \mu_k = 1 \) is locally minimized at \( \mu^* = [\mu_k^*]_{c \times n} \) if and only if

\[
\mu_k^* = \left( \frac{\|x_k - a_i\|^2 - w \ln(p_i)}{\sum_{j=1}^{c} (\|x_k - a_i\|^2 - w \ln(p_j))} \right)^{\frac{1}{m+1}}, \forall i = 1, \ldots, n \text{ and } k = 1, \ldots, n
\]

**Proof.** It is easy to see the equality holds:

\[
\min J_m(\mu, \hat{a}, \hat{p}) = \min \sum_{k=1}^{n} \mu_k^m \left( \|x_k - a_i\|^2 - w \ln(p_i) \right) = \sum_{k=1}^{n} \min \left( \sum_{i=1}^{c} \mu_k^m \left( \|x_k - a_i\|^2 - w \ln(p_i) \right) \right)
\]

subject to \( \sum_{i=1}^{c} \mu_k = 1 \) \( \forall k \)

Since, for all \( k \), the constraints with \( \sum_{i=1}^{c} \mu_k = 1 \) are all the same, i.e. \( g_1(x_1, x_2, \ldots, x_n) = \cdots = g_n(x_1, x_2, \ldots, x_n) \), we may only consider a fixed \( k \) for \( k = 1, \ldots, n \). Thus, \( \forall k \), let the Lagrangian function be

\[
L_2 = \sum_{k=1}^{n} c \mu_k^m \|x_k - a_i\|^2 - w \sum_{k=1}^{n} c \mu_k^m \ln(p_i) + \lambda_k \left( \sum_{i=1}^{c} \mu_k - 1 \right)
\]

where \( \lambda_k \) is a Lagrangian multiplier. With the gradient of \( L_2 \) w.r.t \( \mu_k \) and \( \lambda_k \), we have

\[
\begin{align*}
\frac{\partial}{\partial \mu_k} &= m \mu_k^{m-1} \|x_k - a_i\|^2 - m \mu_k^{m-1} w \ln(p_i) + \lambda_k = 0 \\
\frac{\partial}{\partial \lambda_k} &= \sum_{i=1}^{c} \mu_k - 1 = 0
\end{align*}
\]

\[
\Rightarrow \mu_k = \left( \frac{\lambda_k}{m} \right)^{\frac{1}{m+1}} \left( \|x_k - a_i\|^2 - w \ln(p_i) \right)^{\frac{1}{m+1}}
\]

and \( \left( \frac{\lambda_k}{m} \right)^{\frac{1}{m+1}} = \frac{1}{\sum_{i=1}^{c} (\|x_k - a_i\|^2 - w \ln(p_i))^{\frac{1}{m+1}}} \). This implies that
\[ \mu_k^* = \frac{\| x_k - a_i \|^2 - w \ln(p_i)}{\sum_{i=1}^{r} (\| x_k - a_i \|^2 - w \ln(p_i))^{2+}} \cdot \lambda_k^* = -m \mu_k^{m-1} \left( \| x_k - a_i \|^2 - w \ln(p_i) \right) \]

and the "only if" condition is proved. We next prove the "if" condition by following Theorem 2. If \( a = \hat{a} \) and \( p = \hat{p} \) are fixed, then \( \frac{\partial^2 L_2}{\partial H_k \partial \lambda_k \partial \mu_k} = \frac{\partial^2 L_2}{\partial H_k \partial \lambda_k \partial \mu_k} = 1 \)

Thus, the bordered Hessian w.r.t. \( \mu_k \) and \( \lambda_k \) is

\[
H_k(\mu_k, \lambda_k) = \begin{bmatrix}
0 & 1 & 1 & \cdots & 1 \\
1 & m(m-1)\mu_k^{m-2} \left( \| x_k - a_i \|^2 - w \ln(p_i) \right) & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & \frac{\partial^2 L_2}{\partial \mu_k \partial \lambda_k} \\
\end{bmatrix}
\]

Similar as the proof in Lemma 2, we check all leading principle minors as follows:

\[
|H_3(\mu_k^*, \lambda_k^*)| = \left( \frac{m(m-1)\mu_k^{m-2} \left( \| x_k - a_i \|^2 - w \ln(p_i) \right)}{m(m-1)\mu_k^{m-2} \left( \| x_k - a_i \|^2 - w \ln(p_i) \right)} \right)_{\mu_k = \mu_k^*, \lambda_k = \lambda_k^*} < 0, \ldots \text{ and } \]

Thus, by Theorem 2, \( J_m(\mu, \hat{a}, \hat{p}) \) subject to \( \sum_{i=1}^{r} H_k = 1 \) is locally minimized at \( \mu^* = \mu_k^* \) with \( \mu_k^* = \frac{\left( \| x_k - a_i \|^2 - w \ln(p_i) \right)^{2+}}{\sum_{i=1}^{r} \left( \| x_k - a_i \|^2 - w \ln(p_i) \right)^{2+}} \).

**Lemma 4.** \( J_m \) is continuous on \( M_{jcm} \times (R^r)^{\xi} \times M_p. \)

**Proof.** Since \( \left\{ a_i \rightarrow \| x_k - a_i \|^2 \right\} \), \( \left\{ \mu_k \rightarrow \mu_k^m \right\} \) and \( \left\{ p_i \rightarrow \ln(p_i) \right\} \) are continuous. Thus, the sum of products of \( \left\{ a_i \rightarrow \| x_k - a_i \|^2 \right\} \) and \( \left\{ \mu_k \rightarrow \mu_k^m \right\} \) is continuous and the sum of products of \( \left\{ \mu_k \rightarrow \mu_k^m \right\} \) and \( \left\{ p_i \rightarrow \ln(p_i) \right\} \) is continuous. Therefore, \( J_m \) is continuous on \( M_{jcm} \times (R^r)^{\xi} \times M_p. \)

**Lemma 5.** Let \( \Omega_t \) be the solution set of \( J_m. \) Then, we have that \( J_m(T_m(\mu, a, p)) = J_m(\mu^*, a^*, p^*) < J_m(\mu, a, p) \) for any \( (\mu, a, p) \notin \Omega_t. \)

**Proof.** Let \( (\mu, a, p) \notin \Omega_t. \) Then

\[
J_m(T_m(\mu, a, p)) = J_m(A_2 \circ A_1(\mu, a, p) = J_m(\mu^*, a^*, p^*)
\]

\[<J_m(\mu^*, a^*, p^*) \text{ by Lemma 3} \]

\[<J_m(\mu, a, p) \text{ by Lemma 1} \]

\[<J_m(\mu, a, p) \text{ by Lemma 2.} \]

**Lemma 6.** The BFCM operator \( T_m \) is continuous on \( M_{jcm} \times (R^r)^{\xi} \times M_p. \)
Proof. Since $A_1(\mu, a, p) = E(a, p) = (E_{11}(a, p), E_{12}(a, p), \ldots, E_{cm}(a, p))$ where $E_k(a, p) = \frac{(\|x_k - a\|^2 - w\ln(p_k))^+}{\sum_{j=1}^{n}\|x_j - a\|^2 - w\ln(p_j))} = \mu_{ik}$. Now \(\{a_i \rightarrow \|x_k - a_i\|^2\}, \{p_i \rightarrow \ln(p_i)\}\) and \(\left\{\left(\|x_k - a_i\|^2 - w\ln(p_i]\right) \rightarrow \left(\|x_k - a_i\|^2 - w\ln(p_i)\right)^+\right\}\) are continuous; and the sum of continuous function is continuous. Thus, $E_k(a, p) = \frac{(\|x_k - a_i\|^2 - w\ln(p_i))^{+\theta}}{\sum_{j=1}^{n}(\|x_j - a_i\|^2 - w\ln(p_j))^{+\theta}}$, the quotient of two continuous, is also continuous. Therefore, $E(a, p) = (E_{11}(a, p), E_{12}(a, p), \ldots, E_{cm}(a, p))$ is continuous. Since $A_2(\mu) = (\mu, F(\mu), G(\mu))$, where $F(\mu) = (F_{11}(\mu), F_{12}(\mu), \ldots, F_{cm}(\mu))$ and $G(\mu) = (G_1(\mu), G_2(\mu), \ldots, G_{cm}(\mu))^\top$ with $F_\theta(\mu) = \sum_{k=1}^{n}\|x_k - a\|^2 - w\ln(p_k)^\theta \quad \mu_{ik} = \frac{\sum_{k=1}^{n}\|x_k - a\|^2 - w\ln(p_k)}{\sum_{j=1}^{n}(\|x_j - a\|^2 - w\ln(p_j))}$. Now $\{\mu_{ik} \rightarrow \mu_{ik}\}^\theta$ and $\{\|x_k - \mu_{ik}x_{jk}\|^2\}^\theta$ are continuous, and the sum of continuous functions is also continuous. Thus, $F(\mu) = (F_{11}(\mu), F_{12}(\mu), \ldots, F_{cm}(\mu))$ and $G(\mu) = (G_1(\mu), G_2(\mu), \ldots, G_{cm}(\mu))^\top$ are continuous. Thus $T_m = A_2 \circ A_1$ is continuous on $M_{cm} \times (R^t)^c \times M_p$. □

Lemma 7. Let $[\text{conv} \nu(X)]^c$ be the $c$-fold Cartesian product of the convex hull of $X$, and let $(E(a(0), p(0)), a(0), p(0))$ be the starting point of iteration with $T_m$. Then $(T_m)^c(E(a(0), p(0)), a(0), p(0)) \in M_{cm} \times [\text{conv} \nu(X)]^c \times M_p$, $t = 1, 2, \ldots$, and $M_{cm} \times [\text{conv} \nu(X)]^c \times M_p$ is compact in $M_{cm} \times (R^t)^c \times M_p$.

Proof.

1. Let $(E(a(0), p(0)), a(0), p(0))$ be the starting point of iteration with $T_m$, where $E(a(0), p(0)) = (E_{11}(a(0), p(0)), E_{12}(a(0), p(0)), \ldots, E_{cm}(a(0), p(0)))$ with

\[
E_{ik}(a(0), p(0)) = \frac{\left(\|x_k - a_i(0)\|^2 - w\ln(p_i)\right)^{+\theta}}{\sum_{j=1}^{n}(\|x_j - a_i(0)\|^2 - w\ln(p_j))^{+\theta}} = \mu_{ik}^{(0)}
\]

Then $a_i^{(1)} = F_i(\mu^{(0)}) = \sum_{k=1}^{n}\|x_k - a_i(0)\|^2 - w\ln(p_i) \mu_{ik}^{(0)}$. Let $d_k = \sum_{i=1}^{n}(\mu_{ik}^{(0)})\|x_k - a_{ik}^{(0)}\|$, $\forall k, i$. Thus, $0 \leq d_k \leq 1, \forall k, i$ and $a_i^{(1)} = \sum_{k=1}^{n}d_kx_k$, with $\sum_{k=1}^{n}d_k = \sum_{k=1}^{n}\sum_{i=1}^{n}(\mu_{ik}^{(0)})\|x_k - a_{ik}^{(0)}\| = \sum_{k=1}^{n}(\mu_{ik}^{(0)})\|x_k\|^2 - w\ln(p_i) = 1$. Therefore, $a_i^{(1)} \in \text{conv} \nu(X), \forall i$ and $a(1) \in [\text{conv} \nu(X)]^c$. Continuing recursively, $a^{(t)} \in [\text{conv} \nu(X)]^c \forall t \geq 1$.

Obviously, $p_i^{(1)} = G_i(\mu^{(0)}) = \frac{\sum_{k=1}^{n}(\mu_{ik}^{(0)})\|x_k - a_i^{(1)}\|^2 - w\ln(p_i)}{\sum_{j=1}^{n}(\mu_{jik}^{(0)})\|x_j - a_i^{(1)}\|^2 - w\ln(p_i)} \in M_p$, $i$.\(p^{(1)} \in M_p \forall i \text{ and } \mu^{(1)} \in M_{cm}\) Continuing recursively, $p^{(t)} \in M_p \forall t \geq 1$ and $\mu^{(t)} \in M_{cm} \forall t \geq 1$. Thus, $(T_m)^c(E(a(0), p(0)), a(0), p(0)) \in M_{cm} \times [\text{conv} \nu(X)]^c \times M_p$, $t = 1, 2, \ldots$.

2. Since $X$ is finite, each $x_i \in X$ has finite components. Therefore, the diameter of $X$, that is equal to diameter of $\text{conv} \nu(X)$, is bounded. Since $\text{conv} \nu(X)$ is the convex hull of finitely many generators $x_i$, it is closed. Thus, $\text{conv} \nu(X)$ is bounded and closed in $R^t$, and so $\text{conv} \nu(X)$ is compact. Based on the generalized Heine-Borel theorem, we have that $[\text{conv} \nu(X)]^c$ is also compact.

For claiming that $M_{cm}$ and $M_p$ are compact, let us consider $M_{cm} = \{\mu | \mu_{ik} \in [0, 1] | \sum_{k=1}^{n}1 = 1, \mu_{ik} \in [0, 1] 0 \leq \sum_{k=1}^{n}1\}$ and $M_\rho = \{p = \left[p_{ik}\right]_{k=1}^{n} | \sum_{k=1}^{n}p_k = 1, p_k \in [0, 1]\}$. An argument is similar as that every respect $(M_{cm} = \text{conv}(M_{cm}), M_p = \text{conv}(M_{cm}))$ given in Ball and Hall [1] establishes compactness of $M_{cm}$ and $M_p$. Thus, we have that $M_{cm} \times [\text{conv} \nu(X)]^c \times M_p$ is compact. □
According to Lemmas 4–7 by assessing the condition of Zangwill's convergence theorem, we derived the convergence theorem for the BFCM clustering algorithm.

**Theorem 3.** Let $X = \{x_1, \ldots, x_n\}$ be bounded in $\mathbb{R}^s$ with the BFCM objective function $J_m(\mu, a, p)$ subject to $\sum_{i=1}^{c} \mu_{ik} = 1$ and $\sum_{i=1}^{c} p_i = 1$. Let $T_m : M_{km} \times (\mathbb{R}^s)^c \times M_p \rightarrow M_{km} \times (\mathbb{R}^s)^c \times M_p$ be the BFCM operator as defined in Definition 1. Then, for any BFCM convergent subsequence $\{ (T_m)^{(t)}(E(a^{(0)}, p^{(0)}), a^{(0)}, p^{(0)}) \}$ will tend to the optimal solution $(\mu^*, a^*, p^*)$ in $\Omega_F$, and the BFCM sequence $\{ (T_m)^{(t)}(E(a^{(0)}, p^{(0)}), a^{(0)}, p^{(0)}) \}$ will monotonically converge to the optimal solution $(\mu^*, a^*, p^*)$ in $\Omega_F$.

4. Examples and experimental results

This section presents several examples used to compare the proposed BFCM, BGK and BICS algorithms with the FCM, GK and ICS algorithms. The effectiveness and superiority of the proposed algorithms are indicated by the number of optimal clustering results, error rates, and root mean squared errors (RMSEs) to demonstrate. In all examples, we give $m = 2$, $r = 0.0001$, $\varepsilon = 0.0001$ and $p_i = 1, i = 1, \ldots, c$. 

![Data set images](image-url)

Fig. 12. (a) Four-cluster data set; (b) good clustering results from FCM, ICS, BFCM and BICS; (c) nine-cluster data set; (d) bad clustering results from FCM and ICS; (e) good clustering results from FCM, ICS, BFCM and BICS; (f) sixteen-cluster data set; (g) bad clustering results from FCM and ICS; and (h) good clustering results from FCM, ICS, BFCM and BICS.
Example 3. In this example, we assessed three data sets drawn from uniform distributions. The first data set comprised four clusters and each cluster contained 100 data points (Fig. 12(a)). The second data set comprised nine clusters and each cluster contained 100 data points (Fig. 12(c)). The third data set contained 16 clusters and each cluster comprised 100 data points (Fig. 12(f)). We used 100 randomly generated initializations assignments for the FCM and ICS algorithms. Regarding the data set illustrated in Fig. 12(a), the FCM and ICS algorithms obtain 100 times to cluster the data set with optimal clustering results (Fig. 12(b)). Regarding the data set depicted in Fig. 12(c), the FCM and ICS algorithms obtain approximately 75 times to cluster the data set with optimal clustering results like those shown in Fig. 12(e). Regarding the data set illustrated in Fig. 12(f), the FCM and ICS algorithms only obtain 16 times and 13 times to cluster the data set with optimal clustering results like those depicted in Fig. 12(h). We also discovered that the number of poor clustering results (Fig. 12(d) and (g)) increased when the cluster number increased. Specifically, the clustering results of the FCM and ICS algorithms actually depend on initializations when the cluster number of the data set increases. We executed the proposed BFCM and BICS algorithms for the data sets illustrated in Fig. 12(a), (c), and (f). We also used 100 randomly generated initialization assignments for these algorithms. Regarding the data set depicted in Fig. 12(a), the BFCM and BICS algorithms obtain 100 times to cluster the data set with optimal clustering results like those illustrated in Fig. 12(b). Regarding the data set shown in Fig. 12(c), the BFCM and BICS algorithms obtain 100 times to cluster the data set with optimal clustering results like those depicted in Fig. 12(e). Regarding the data set illustrated in Fig. 12(f), the BFCM and BICS algorithms obtain 96 and 80 times, respectively, to cluster the data set with optimal clustering results like those shown in Fig. 12(h). We determined that the BFCM and BICS algorithms were actually more robust against initializations than the FCM and ICS were.

Example 4. In this example, we used a data set with a sample size of 1000 (Fig. 13(a)) and generated from a two-component Gaussian mixture distribution

$$f(x; \varpi, \theta) = \sum_{k=1}^{c} \varpi_k f(x; \theta_k) = \sum_{k=1}^{c} \varpi_k (2\pi)^{-\frac{d}{2}} |\Sigma_k|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

where the parameters are $\varpi_1 = \varpi_2 = 0.5$, $\mu_1 = (0, 0)^T$, $\mu_2 = (0, 5)^T$ and $\Sigma_1 = \Sigma_2 = \begin{pmatrix} 5 & -4.5 \\ -4.5 & 5 \end{pmatrix}$. When we used 100 randomly generated initialization assignments for the FCM and BFCM algorithms, these algorithms obtain 100 times to classify the data set with poor clustering results like those illustrated in Fig. 13(b). When we used 100 randomly generated initialization assignments for the GK and BGK algorithms, these algorithms obtain 100 times to classify the data set with optimal clustering results like those depicted in Fig. 13(c). These results indicate that the GK and BGK algorithms demonstrate more favorable clustering results than the FCM and BFCM algorithms do.

Furthermore, we assessed another data set with a sample size of 1000 (Fig. 14(a)) and generated from a two-component Gaussian mixture distribution with parameters $\varpi_1 = \varpi_2 = 0.5$, $\mu_1 = \mu_2 = (0, 0)^T$, $\Sigma_1 = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$ and $\Sigma_2 = \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}$. When we used 100 randomly generated initialization assignments for the BGK algorithm, it obtained only 28 times to classify the data set with optimal clustering results like those shown in Fig. 14(c). When we used 100 randomly generated initialization assignments for the BGK algorithm, it obtained 100 times to classify the data set with optimal clustering results like those illustrated in Fig. 14(c). These results indicate that the BGK algorithm is more robust against initializations than the GK algorithm is. Furthermore, we also assessed the accuracy of the clustering algorithms by using the RMSE criterion. The RMSEs evaluated for the data sets depicted in Figs. 13(a) and 14(a) are shown in Table 3. Generally, the BGK clustering algorithm is the most effective among the algorithms.

Although the true number of clusters for the data set illustrated in Fig. 14 is 2 (because this data set was generated from a two-component Gaussian mixture distribution), one may consider that the true number of clusters is also possibly 4 or 5,
Table 3
RMSEs of $\mu_1 = (\mu_{11}, \mu_{12})$ and $\mu_2 = (\mu_{21}, \mu_{22})$ from different algorithms with 100 randomly initialization assignments for Figs. 10(a) and 11(a).

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>RMSEs of $\mu_1$ for Fig. 10(a)</th>
<th>RMSEs of $\mu_2$ for Fig. 10(a)</th>
<th>RMSEs of $\mu_1$ for Fig. 11(a)</th>
<th>RMSEs of $\mu_2$ for Fig. 11(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCM</td>
<td>(1.1865, 0.4154)</td>
<td>(1.1052, 0.1623)</td>
<td>(0.8051, 1.1323)</td>
<td>(0.8013, 1.1559)</td>
</tr>
<tr>
<td>BFCM</td>
<td>(1.1214, 0.2705)</td>
<td>(1.1418, 0.3146)</td>
<td>(0.8027, 1.1450)</td>
<td>(0.8045, 1.1331)</td>
</tr>
<tr>
<td>GK</td>
<td>(0.0181, 0.0425)</td>
<td>(0.0207, 0.0233)</td>
<td>(0.7906, 0.9400)</td>
<td>(0.8135, 0.9638)</td>
</tr>
<tr>
<td>BGK</td>
<td>(0.0164, 0.0491)</td>
<td>(0.0219, 0.0186)</td>
<td>(0.2301, 0.2568)</td>
<td>(0.2408, 0.2832)</td>
</tr>
</tbody>
</table>

Fig. 14. (a) Two-cluster data set; (b) bad clustering results from GK; and (c) good clustering results from GK and BGK.

Fig. 15. (a) Two-cluster data set and (b) good clustering results.

Table 4
Error rates and RMSEs of $\mu_1 = (\mu_{11}, \mu_{12})$ and $\mu_2 = (\mu_{21}, \mu_{22})$ from GK and BGK with 100 randomly initialization assignments for Fig. 12(a).

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Error rate</th>
<th>RMSEs of $\mu_1$</th>
<th>RMSEs of $\mu_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GK</td>
<td>0.3406</td>
<td>(0.3923, 0.0020)</td>
<td>(0.4591, 0.0020)</td>
</tr>
<tr>
<td>BGK</td>
<td>0.0170</td>
<td>(0.0210, 0.0032)</td>
<td>(0.0205, 0.0031)</td>
</tr>
</tbody>
</table>

according to human vision. This is because a larger covariance matrix was used. To evaluate an exact 2-cluster data set, we used another data set. We assessed a data set with two spread lines and each spread line had a sample size of 500 (Fig. 15); this data set was generated from a two-component Gaussian distribution with the parameters $\mu_1 = \mu_2 = 0.5, \Sigma_1 = \Sigma_2 = [0.5, 0; 0, 0.00001]$, where each component has a sample size 500. We then rotated one component at 9$^\circ$ (i.e. $\frac{\pi}{18}$) and the other component at $-9^\circ$ (i.e. $-\frac{\pi}{18}$); therefore, the data set demonstrated two line-shaped clusters that intersected at (0,0). We executed the GK and BGK algorithms for this data set. When we used 100 randomly generated initialization
assignments for the GK algorithm, it obtained only 32 times to classify the data set with optimal clustering results like those shown in Fig. 15(b). When we used 100 randomly generated initialization assignments for the BGK algorithm, it obtained 94 times to classify the data set with optimal clustering results like those depicted in Fig. 15(b). These results indicate that the BGK algorithm is more robust against initializations than the GK algorithm is. Furthermore, we also evaluated the accuracy of the clustering algorithms by using the RMSE criterion. These RMSEs evaluated for the data sets shown in Fig. 15(a) are shown in Table 4. As shown in this table, the BGK algorithm is generally more excellent than the GK algorithm.

**Example 5.** We used a simulated data set (Fig. 16(a)) in this example. This data set comprised 16 clusters and each cluster contained 100 data points. These 16 clusters are clearly separated. However, when we used the FCM algorithm with the cluster number \( c = 16 \), it returned only 23 optimal clustering results. When we used the BFCM and GK algorithms with the cluster number \( c = 16 \), they returned 77 optimal clustering results. When we used the BGK algorithm with the cluster number \( c = 16 \), it returned 91 optimal clustering results. These results indicate that the BFCM algorithm is more favorable than the FCM algorithm; however, the BGK algorithm is the most excellent algorithm among them.

**Example 6.** The data set (Fig. 17(a)) was generated from a two-dimensional five-component Gaussian mixture distribution with two disk-like shapes and three oval shapes. We executed the FCM, BFCM, GK, and BGK algorithms for this data set and assigned 100 randomly generated initialization assignments; we discovered that both the FCM and BFCM algorithms returned poor clustering results (Fig. 17(b)–(d)). The GK algorithm returned 52 poor clustering results (Fig. 17(e)–(f)) and 48 optimal clustering results (Fig. 17(g)). However, the BGK algorithm returned 100 optimal clustering results (Fig. 17(h)), indicating that it has the most favorable clustering results among the algorithms.

**Example 7** (*banana-shaped data*). In this example, we assessed a data set with four clusters with banana-shaped distributions (Fig. 18). We executed the FCM, BFCM, GK, and BGK algorithms for this data set. Table 5 shows the average error rates of these algorithms. The error rates of the FCM, BFCM, GK, and BGK algorithms were 10.78%, 12.34%, 4.38%, and 0%, respectively; the BGK algorithm registered the lowest error rate among the algorithms. Thus, the BGK algorithm is more effective than the FCM, BFCM, and GK algorithms in analyzing banana-shaped data; however, the effectiveness of the BFCM algorithm is not higher than that of the FCM algorithm.
Fig. 17. (a) Data set; (b)–(d) bad clustering results from FCM and IFCM; (e)–(f) bad clustering results from GK; (g) good clustering results from GK and BGK; and (h) identified 5 clusters.

Fig. 18. (a) Banana-shaped data set and (b) good clustering results for BGK.
Example 8 (Iris data). The Iris data set is a typical test data set for most classification techniques in pattern recognition. The data set comprises 50 samples from each of three species of iris flowers (Iris setosa, Iris virginica, and Iris versicolor). The attributes of Iris data were measured with the length and the width of the sepal and petal. Fig. 19 illustrates the scatterplot of Iris data (see also Wikipedia for more details on Iris data). As shown in Fig. 19, each plot comprises two clusters and the larger cluster contains two overlapped clusters. We executed the FCM, BFCM, GK, and BGK algorithms for the Iris data set. We used 100 times of random initializations for executing these algorithms and recorded the misclassifications of each method. Table 6 shows the average error rates (error numbers) of the FCM, BFCM, GK, and BGK algorithms, indicating that the error rates of the GK and BGK algorithms are 10%, which are lower than those of the other algorithms. Thus, the GK and BGK algorithms are more effective than the FCM and BFCM algorithms for the Iris data set. However, in our simulation, we determined that the FCM and GK algorithms were not affected by initializations; therefore, our bias-correction term did not reduce the error rate for the Iris data set.

Example 9 (Crab data set). In this example, we applied the proposed BFCM and BGK clustering algorithms to a real data set, crab data of Campbell and Mahon [3]. We assessed a data set of crab (belonging to the genus Leptograpsus) observations, and this data set comprised measures over a sample of 100 blue crabs; Fig. 20(a) depicts this data set and “+” denotes the cluster of male crabs and ”•“ denotes the cluster of female crabs. Five measurements were conducted for each specimen: the width of the front lip, rear width, length along the midline, maximum width of the carapace, and body depth in millimeters. They were classified into two clusters, namely male crabs and female crabs, and each cluster comprised 100 crabs. Fig. 20(a) illustrates the scatter plot of the second and third features of the data set. Peel and MacLaren [19] fitted this data set by using both normal and $t$ distributions with equal covariance matrices in which the two models resulted in approximately the same
error rate with 0.18–0.19. We executed the BFCM and BGK algorithms for this data set with 100 randomly generated initialization assignments; Fig. 20(b) depicts the clustering result of the BFCM algorithm, indicating that the error rate is 0.42, and Fig. 20(c) illustrates the clustering result of the BGK algorithm, indicating that the error rate is only 0.09. We determined that the BFCM algorithm registered a higher error rate than those of the normal and $t$ distribution clustering algorithms proposed by Peel and McLachlan [19]. However, the error rate of the BGK algorithm was lower than those of the normal and $t$ distribution clustering algorithms.

Example 10 (Flea beetle data set). We also applied the BFCM and BGK algorithms to a data set of flea beetle observations reported by Lubischew [13]; Fig. 21(a) shows this data set, where “1” denotes the cluster of the heikertingeri, “2” denotes the cluster of the concinna species, and “3” denotes the cluster of the heptapotamica species. The data set comprised 74 data points with three species: concinna (21), heikertingeri (31), and heptapotamica (22). Each data point was obtained by measuring two characteristics of a beetle: the maximal width of the aedeagus in the fore-part in microns and the front angle of the aedeagus (1 unit = $75$ microns). We executed the BFCM and BGK algorithms for this data set with 100 randomly generated initialization assignments. Fig. 21(b) illustrates the clustering result of the BFCM algorithm, indicating an error rate of 0.1486. Fig. 21(c) depicts the clustering result of the BGK algorithm, indicating an error rate of 0.0135. These results show that the proposed BGK performs effectively for the real data set of flea beetles as indicated by its low error rate (only 0.0135), which is considerably lower than that of the BFCM clustering algorithm.

5. Conclusion and discussion

Initializations usually influence the performance of the FCM algorithm and its generalizations. In this paper, we proposed bias-correction fuzzy clustering algorithms. The bias-correction term is similar to entropy and it can present the total information for fuzzy c-partitions that can be used for adjusting the effects of poor initializations to most fuzzy clustering
algorithms. During the algorithm implementations, we also used an updating equation for a weighted parameter to ensure that the bias correction decreases when the number of iteration increases. This updating equation can stabilize algorithms when the number of iterations increases gradually. Several numerical examples and real data sets demonstrate the effectiveness of the proposed bias-correction fuzzy clustering algorithms.

The k-means algorithm is usually affected by initializations. We may use the same bias-correction idea to construct a bias-correction k-means algorithm to enable it to become robust against initializations. Furthermore, the PCM clustering algorithm is evaluated by relaxing the FCM algorithm constraint that specifies that the memberships of a data point across classes sum to 1. Because the PCM algorithm has a tendency to produce coincident clusters, it can serve as an efficient mode seeking algorithm if it has optimal initializations. However, the performance of the PCM algorithm heavily depends on the assigned initializations. In this sense, we may also use the bias-correction idea to construct a bias-correction PCM so that it can be robust against initializations and become an extremely effective seeking algorithm. These will be our future research topics.

Acknowledgements

The authors would like to thank the anonymous referees for their helpful comments in improving the presentation of this paper. This work was supported in part by the National Science Council of Taiwan, under Grant NSC101-2118-M-033-002-MY2.

References