A ROBUST FUZZY CLASSIFICATION MAXIMUM LIKELIHOOD CLUSTERING FRAMEWORK

MIIN-SHEN YANG*,‡, CHIH-YING LIN* and YI-CHENG TIAN*,†

*Department of Applied Mathematics
Chung Yuan Christian University
Chung-Li 32023, Taiwan
Department of Applied Mathematics,
†Center for General Education,
Hsin Sheng College of Medical Care and Management,
Chung-Li, Taiwan
‡msyang@math.cycu.edu.tw

Received 16 October 2012
Revised 26 April 2013

In 1993, Yang first extended the classification maximum likelihood (CML) to a so-called fuzzy CML, by combining fuzzy c-partitions with the CML function. Fuzzy c-partitions are generally an extension of hard c-partitions. It was claimed that this was more robust. However, the fuzzy CML still lacks some robustness as a clustering algorithm, such as its inability to detect different volumes of clusters, its heavy dependence on parameter initializations and the necessity to provide an a priori cluster number. In this paper, we construct a robust fuzzy CML clustering framework that has a robust clustering method. The eigenvalue decomposition of a covariance matrix is firstly considered using the fuzzy CML model. The Bayesian information criterion (BIC) is then used for model selection, in order to choose the best model with the optimal number of clusters. Therefore, the proposed robust fuzzy CML clustering framework exhibits clustering characteristics that are robust in terms of the parameter initialization, robust in terms of the cluster number and also in terms of its capability to detect different volumes of clusters. Numerical examples and real data applications with comparisons are provided, which demonstrate the effectiveness and superiority of the proposed method.

Keywords: Clustering; model-based clustering; k-means; fuzzy clustering; fuzzy c-means; classification maximum likelihood; robustness.

1. Introduction

Data analysis is the science of analyzing real world data. Cluster analysis is a useful tool for data analysis. It identifies clusters as the groups of data points in a data set with the greatest similarity within the same cluster and the largest dissimilarity between different clusters. Cluster analysis is a branch of statistical multivariate analysis. In contrast, learning and recognition mostly begin with clustering, so cluster analysis represents an unsupervised learning in pattern recognition. From the statistical point of view,
clustering methods can be generally categorized as either having a (probability) model-based approach or a nonparametric approach. A model-based clustering approach assumes that the data set obeys a mixture model of probability distributions, so a mixture likelihood approach to clustering is used.\(^5,6\) For a nonparametric approach, clustering methods are generally based on an objective function of similarity or dissimilarity measures and partitional methods are popularly used.\(^2,4\) Most partitional methods suppose that the data set can be represented by finite cluster prototypes with their own objective functions. Therefore, the definition of the dissimilarity (or distance) between points and cluster prototypes is essential for partitional methods. The most popular partitional methods for cluster prototypes are k-means,\(^7,8\) trimmed k-means,\(^9,10\) fuzzy c-means (FCM),\(^11,13\) mean shift,\(^14,16\) and possibilistic c-means (PCM).\(^17,19\)

There are two categories of model-based clustering algorithms. One is based on the expectation and maximization (EM) algorithm,\(^20,21\) and the other is based on the classification maximum likelihood (CML).\(^22,24\) The EM algorithm is most popular, but it is sensitive to initial values and to cluster shapes in which an a priori number of components (clusters) must be given.\(^21\) Scott and Symons\(^23\) first proposed the CML procedure. Yang\(^25\) then extended the original CML to the so-called fuzzy CML (FCML) based on fuzzy c-partitions. Banfield and Raftery\(^26\) extended the original CML by using the eigenvalue decomposition of a covariance matrix to allow the specification of which feature (orientation, size and shape) is common to all clusters and which differs between clusters. Fraley and Raftery\(^27,28\) continued to develop a more general model-based clustering procedure.

In the literature, there are some applications of the fuzzy CML.\(^29-33\) However, the FCML is not sufficiently robust in clustering, so it cannot detect different volumes (sizes and shapes) of clusters and also requires an a priori number of clusters with parameter initialization. In this paper, a robust framework is constructed for FCML clustering. It is firstly necessary to consider the eigenvalue decomposition of a covariance matrix in the fuzzy CML model. The Bayesian information criterion (BIC) is then used to choose the best model with the optimal number of clusters. The remainder of this paper is organized as follows. Section 2 provides a review of the FCML clustering method proposed by Yang.\(^25\) In Sec. 3, the robust FCML clustering framework is constructed. Section 4 contains some numerical comparisons of the proposed method with existing methods. The method is then applied to real data sets. Finally, conclusions are stated in Sec. 5.

\textbf{2. Fuzzy Classification Maximum Likelihood Clustering}

The classification maximum likelihood (CML)\(^22,24\) is a classification approach that approximates the maximum likelihood estimates. Let \(X = \{x_1, \ldots, x_n\}\) be a random sample of size, \(n\), from a population that consists of \(c\) different subpopulations where \(f_k(x; \theta)\) is the density function of the \(k\)th subpopulation with an unknown vector of parameters, \(\theta\), and let \(P = (P_1, \ldots, P_c)\) be a partition of \(X\). Then, the joint density of \(X\) is
\[
\prod_{i=1}^{n} \prod_{k \in P_i} f_i(x_i; \theta_k). \]
In the CML method, \( P = (P_1, \ldots, P_c) \) and \( \theta = (\theta_1, \ldots, \theta_c) \) are estimated by maximizing the log likelihood function \( \sum_{i=1}^{n} \sum_{k \in P_i} \ln f_i(x_i; \theta_k) \), so the CML objective function \( L_{CML} \) for the grouped data \( X = \{x_1, \ldots, x_n\} \) is
\[
L_{CML}(P, \theta; X) = \sum_{i=1}^{n} \sum_{k \in P_i} \ln f_i(x_i; \theta_k).
\]

If a \( d \)-variate Gaussian mixture model with mean vectors, \( \alpha_k \), and covariance matrices, \( \Sigma_k \), is considered, then the CML objective function, \( L_{CML} \), becomes
\[
L_{CML}^d(P, \alpha, \Sigma; X) = -\frac{nd}{2} \ln(2\pi) - \frac{1}{2} \sum_{k=1}^{c} n_k \ln |\Sigma_k| - \frac{1}{2} \sum_{k=1}^{c} \sum_{i \in P_k} (x_i - \alpha_k) \Sigma_k^{-1} (x_i - \alpha_k)
\]
where \( n_k = |P_k| \) is the cardinality of \( P_k \). If the sample cross-product matrix, \( W_k \), for the \( k \)th cluster is \( W_k = \sum_{i \in P_k} (x_i - \overline{x}_k)(x_i - \overline{x}_k)' \), then
\[
\hat{P} = \arg \max_{P} L_{CML}(P, \hat{\alpha}, \hat{\Sigma}; X) = \arg \min_{P} \left[ \sum_{k=1}^{c} n_k \ln \left| \frac{W_k}{n_k} \right| \right].
\]

This criterion is the CML originally proposed by Scott and Symons.23

Fuzzy set theory, as proposed by Zadeh,34 has been widely used in cluster analysis since Ruspini35 introduced the notion of fuzzy c-partitions (see also Refs. 11−13). Yang25 made the fuzzy extension of the original CML to the fuzzy CML (FCML), based on fuzzy c-partitions. This section firstly introduces the concept of hard and fuzzy c-partitions and then the FCML method.

The CML objective function is \( L_{CML}(P, \theta; X) = \sum_{k=1}^{c} \sum_{i \in P_k} \ln f_i(x_i; \theta_k) \). The parameter, \( \alpha_k > 0 \), is the proportion of the \( k \)th subpopulation in the whole population with the constraint, \( \sum_{k=1}^{c} \alpha_k = 1 \). The indicator functions, \( z_1, \ldots, z_c \), are used with \( z_j(x) = 1 \) if \( x \in P_j \) and \( z_j(x) = 0 \) if \( x \notin P_j \) for all \( x \) in \( X \) and for all \( j = 1, \ldots, c \). This is generally known as clustering \( X \) into \( c \) clusters using \( z = (z_1, \ldots, z_c) \) and is called the hard c-partition of \( X \). Therefore, the CML objective function becomes
\[
L_{CML}(P, \alpha, \theta; X) = \sum_{j=1}^{c} \sum_{i \in P_j} \ln \alpha_j f_j(x_i; \theta) = \sum_{j=1}^{c} \sum_{i \in P_j} z_j(x_i) \ln \alpha_j f_j(x_i; \theta).
\]
For the extension that allows the indicator functions, \( z_j(x_i) \), to be functions (known as membership functions), the values are assumed to be in the interval \([0, 1]\), such that \( \sum_{j=1}^{c} z_j(x_i) = 1 \) for all \( x_i \in X \). In this case, the data set, \( X \), is said to have a fuzzy c-partition, \( z = (z_1, \ldots, z_c) \), according to basic idea of fuzzy sets, as proposed by Zadeh. Based on this fuzzy extension, Yang proposed the FCML objective function as follows:

\[
J_{FCML}(z, \alpha, \theta; X) = \sum_{j=1}^{c} \sum_{i=1}^{m} z_{ij}^m \ln f_j(x_i; \theta) + r \sum_{j=1}^{c} \sum_{i=1}^{m} z_{ij}^m \ln \alpha_j
\]

subject to \( \sum_{j=1}^{c} \alpha_j = 1, \alpha_j > 0 \) and \( \sum_{j=1}^{c} z_{ij} = 1 \) for all \( i = 1, \ldots, n \) for all \( z_{ij} \in [0,1] \). The criterion, \( m > 1 \), represents the degree of fuzziness and \( r \geq 0 \) is a scale parameter for the penalty term \( \sum_{j=1}^{c} \sum_{i=1}^{m} z_{ij}^m \ln \alpha_j \). Therefore, the FCML procedure is created by choosing a fuzzy c-partition, \( z \), and a proportion, \( \alpha \), and an estimate, \( \theta \), that maximize \( J_{FCML}(z, \alpha, \theta; X) \). Yang used a multivariate normal distribution with an identity covariance matrix and then derived a penalized fuzzy c-means clustering (PFCM) algorithm, where the PFCM objective function is exactly the FCM objective function with the addition of the penalty term \( \sum_{j=1}^{c} \sum_{i=1}^{m} z_{ij}^m \ln \alpha_j \). There are several applications of the FCML, but the FCML algorithm still lacks some robustness, in terms of its inability to detect different sizes and shapes of clusters, its heavy dependence on parameter initializations and also the necessity to provide an a priori number of clusters. In the next section, a robust framework for FCML clustering is constructed.

3. The Proposed robust FCML Clustering Framework

Although the CML is a probability model-based cluster analysis that is used to understand some existing clustering methods, such as agglomerative hierarchical clustering, k-means and the criterion of Friedman and Rubin, the CML does not allow the specification of which feature (orientation, size and shape) is common to all clusters and which may differ between clusters. In order to address these drawbacks of the CML, Banfield and Raftery first created a so-called model-based clustering by extending the CML with the eigenvalue decomposition of the covariance matrix. Banfield and Raftery’s eigenvalue decomposition approach is firstly narrated.

Since the CML originally proposed by Scott and Symons assumes that all the components (clusters) are different, with a Gaussian mixture model, where

\[
\hat{P} = \arg \max_{P} L_{CML}(P, \hat{\Sigma}, \hat{\Sigma}; X) = \arg \min_{P} \left\{ \sum_{j=1}^{c} \frac{n_k}{n} \ln \left( \frac{W_k}{n_k} \right) \right\}, \text{ then in order to allow the specification of which feature is common to all clusters and which may differ between clusters, Banfield and Raftery re-parameterized the covariance matrix, } \Sigma_k, \text{ in terms of its eigenvalue decomposition with}
\]
\[ \Sigma_k = D_k \Lambda_k D_k^T = \lambda_k D_k \Lambda_k D_k^T \]

where \( \lambda_k = \max \{ \text{diag}(A_k) \} \) determines the size of the \( k \)th cluster, \( D_k \) is an orthogonal matrix representing its orientation, and \( A_k = \text{diag}(a_{1,k}, \ldots, a_{w,k}) \) represents its shape. If different constraints for its eigenvalue decomposition are given, the criterion for partition, \( P \), in the CML becomes the following clustering criteria:

1. When \( \lambda_k = \sigma^2 k \), \( A_k = I \) and \( D_k = I \) for all \( k = 1, \ldots, c \) and the clustering is spherical and the same size as for \( \hat{P} = \arg\min_P \text{tr} (\sum_{k=1}^c W_k) = \arg\min_P \sum_{k=1}^c \sum_{x \in P_k} \| x - \hat{\mu}_k \|^2 \). This criterion was originally proposed by Ward.\(^{16}\)

2. When \( \lambda_k = \sigma^2 k \), \( A_k = A \) and \( D_k = D \) for all \( k = 1, \ldots, c \) and the clustering is ellipsoidal and the same size, shape and orientation as for \( \hat{P} = \arg\min_P \sum_{k=1}^c W_k \). This criterion was originally proposed by Friedman and Rubin.\(^{37}\)

3. When there is no constrain to \( \Sigma_k \) for all \( k = 1, \ldots, c \), the clustering is ellipsoidal and different in size, shape and orientation to that for \( \hat{P} = \arg\min_P \left\{ \sum_{k=1}^c n_k \ln \frac{W_k}{n_k} \right\} \). This criterion was the original type proposed by Scott and Symons.\(^{23}\)

4. When \( \lambda_k = \sigma^2 k \), \( A_k = I \) and \( D_k = I \) for all \( k = 1, \ldots, c \) and the clustering is spherical and different in size to that for \( \hat{P} = \arg\min_P \left\{ \sum_{k=1}^c n_k \ln \left( \text{tr} \left( \frac{W_k}{n_k} \right) \right) \right\} \). This criterion was proposed by Banfield and Raftery.\(^{26}\)

5. When there is only the constraint, \( A_k = A \), where \( A \) is known, for all \( k = 1, \ldots, c \), each cluster is different in size and orientation but has the same known shape, so \( \hat{P} = \arg\min_P \left\{ \sum_{k=1}^c n_k \ln \left( \text{tr} \left( A_k^{-1} \Omega_k \right) \right) \right\} \). This criterion was also proposed by Banfield and Raftery.\(^{26}\)

For \( X = \{ x_1, \ldots, x_n \} \) as a random sample of size, \( n \), from a mixture of probability density functions, \( f(x; \alpha, \theta) \), and the fuzzy \( c \)-partition, \( z = (z_1, \ldots, z_n) \), with values in the interval \([0, 1]\), such that \( \sum_{j=1}^c z_{ij}(x_i) = 1 \) for all \( x_i \in X \), Yang\(^{16}\) proposed the FCML objective function as follows:

\[ J_{\text{FCML}}(z; \alpha, \theta; X) = \sum_{j=1}^c \sum_{i=1}^n z_{ij}^a \ln f_j(x_i; \theta) + \sum_{j=1}^c \frac{m}{n} \sum_{i=1}^n z_{ij}^a \ln \alpha_j \]
subject to $\sum_{j=1}^{c} \alpha_j = 1$, $\alpha_j > 0$ and $\sum_{i=1}^{n} z_{ij} = 1$ for $i = 1, \ldots, n$ for all $z_{ij} \in [0,1]$. The eigenvalue decomposition concept is used to advance the FCML model by reparameterizing the covariance matrix, $\Sigma_j$, in terms of its eigenvalue decomposition with $\Sigma_j = D_j A_j D_j'$. The distribution, $f_j(x; \theta)$, is a multivariate normal distribution. The decomposition of $\Sigma_j = \lambda_j D_j A_j D_j'$ produces the following five cases.

**Case 1:** If the parameters of $\Sigma_j$ are allowed to vary, then

$$J_{\text{FCML}}(z, \alpha, u, \Sigma) = \sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij} \ln\left(\frac{1}{(2\pi)^{n/2} \det(\Sigma_j)^{1/2}} \exp\left(-\frac{(x_i - u_j)' \Sigma_j^{-1} (x_i - u_j)}{2}\right) \right) + \sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij} \ln \alpha_j.$$ If $d_j = \left(-\frac{1}{2}\right) \ln|\Sigma_j| + (x_i - u_j)' \Sigma_j^{-1} (x_i - u_j) + d \ln(2\pi)) + r \ln \alpha_j$, then the lagrangian $\tilde{J}_{\text{FCML}}$ of $J_{\text{FCML}}$ is $\tilde{J}_{\text{FCML}}(z, \alpha, u, \Sigma, \lambda) = \sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij} d_i - \ell_j \left(\sum_{j=1}^{c} \alpha_j - 1\right) - \ell_z \left(\sum_{j=1}^{c} z_{ij} - 1\right)$. The first derivatives of the lagrangian, $\tilde{J}_{\text{FCML}}$, are taken with respect to the parameters, $z_{ij}, \alpha_j, u_j, \Sigma_j$, and these are set to 0. Therefore, the necessary conditions for the maximization $(z, \alpha, u, \Sigma)$ of $J_{\text{FCML}}(z, \alpha, u, \Sigma)$ can be derived as follows:

$$\alpha_j = \frac{\sum_{i=1}^{n} z_{ij}^n}{\sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij}^n}, \ j = 1, \ldots, c \quad (1)$$

$$u_j = \frac{\sum_{i=1}^{n} z_{ij} x_i}{\sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij}^n}, \ j = 1, \ldots, c \quad (2)$$

$$z_{ij} = \left\{ \begin{array}{ll}
\frac{c}{i} \left( \frac{d_{ij}}{d_j} \right)^{1/i} & , \ j = 1, \ldots, c , \ i = 1, \ldots, n \\
\sum_{i=1}^{n} z_{ij}^n (x_i - u_j)' (x_i - u_j) & , \ j = 1, \ldots, c \end{array} \right. \quad (3)$$

$$\Sigma_j = \sum_{i=1}^{n} z_{ij}^n (x_i - u_j)' (x_i - u_j)' \sum_{i=1}^{n} z_{ij}^n , \ j = 1, \ldots, c \quad (4)$$

The factor, $r$, of the penalty term, $\sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij}^n \ln \alpha_j$, in the FCML objective function, $J_{\text{FCML}}$, is now considered. Since the penalty term is a function of the mixing proportions, $\alpha_j$, the factor, $r$, must be proportional to the impact of the mixing proportions, $\alpha_j$, on $J_{\text{FCML}}$. In fact, when $r = 0$, the penalty term, $\sum_{j=1}^{c} \sum_{i=1}^{n} z_{ij}^n \ln \alpha_j$, has no impact on $J_{\text{FCML}}$, but
when \( r = 1 \), the mixing proportions, \( \alpha_j \in [0, 1] \), with \( \sum_{j=1}^{J} \alpha_j = 1 \), have a full impact on \( J_{FCML} \). In this sense, a constraint for \( r \) can be set between 0 and 1, i.e. \( r \in [0,1] \). A way to estimate the factor, \( r \), is then proposed. Since the true values of the mixing proportions, \( \alpha_j \), depend on the cluster structure of the data set, the value of \( r \) can be estimated according to its data structure. Because \( r \in [0,1] \), the correlation coefficients between data attributes of the data set can be used to estimate \( r \), using the following process. Let \( X = \{x_1, \ldots, x_n\} \) be a data set with \( n \) observations and \( d \) attributes. We consider \( y_s = (x_{s1}, \ldots, x_{sd}) \), \( s = 1, \ldots, d \) as the data vector of \( s \)th attribute for the data set, \( X = \{x_1, \ldots, x_n\} \). The correlation coefficients between these attribute vectors with \( \rho_{s \neq t} = \frac{\sum_{i=1}^{n}(y_{si} - \bar{y}_s)(y_{ti} - \bar{y}_t)}{\sqrt{\sum_{i=1}^{n}(y_{si} - \bar{y}_s)^2} \sqrt{\sum_{i=1}^{n}(y_{ti} - \bar{y}_t)^2}}, s \neq t \) are then calculated. The initial value of \( r \) is 1/2 and the value is increased for positive correlation coefficients and decreased for negative correlation coefficients. Therefore, the following estimate is proposed for \( r \):

\[
 r = \frac{1+\sum_{s \neq t} \sum_{i=1}^{n} \rho_{ys,it}}{2} \tag{5}
\]

Since the value of \( r \) may exceed 1, if \( r \geq 1 \), we set \( r = 1 \). In contrast, the value of \( r \) may become less than 0, so if \( r \leq 0 \), we set \( r = 0 \). In this case, the factor, \( r \), is an aggregate of the degrees of correlation between data attributes for the given data set. If the data set has a larger aggregated degree of correlation between attribute vectors, then the mixing proportions, \( \alpha_j \), have a larger impact on the FCML objective function, \( J_{FCML} \).

**Case 2:** If the covariance matrix is replaced with \( \Sigma_j = \lambda I \), then all clusters are the same size. In this case, the objective function becomes the penalized FCM. The updated Eq. (4) is replaced with the following equation:

\[
 \lambda = \frac{\sum_{i=1}^{n} \sum_{j=1}^{J} z_{ij} tr((x_i - u_j)(x_i - u_j)')} {d \sum_{i=1}^{n} z_{ii}^{\alpha}} \tag{6}
\]

**Case 3:** If the cluster sizes are changed to \( \Sigma_j = \lambda_j I \), the updated Eq. (4) is replaced with the following equation:

\[
 \lambda_j = \frac{\sum_{i=1}^{n} z_{ij}^{\alpha} tr((x_j - u_j)(x_j - u_j)')} {d \sum_{i=1}^{n} z_{ij}^{\alpha}} \tag{7}
\]
Case 4: All clusters are the same size and shape, but have different orientations, so the covariance matrix can be defined in the form, $\Sigma_j = \lambda_j D_j A^T D_j$. Let $W_j$ be the sample cross-product matrix for the kth cluster, written as $W_j = \sum_{i=1}^n z_i (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T$. Note that $W_j / n_j$ is the MLE of $\Sigma_j$. The eigenvalue decomposition of $W_j$ becomes $W_j = L_j \Omega_j L_j^T$, where $\Omega_j = \text{diag}\{w_{1j}, \ldots, w_{mj}\}$ and $w_{mj}$ is the mth eigenvalue of $W_j$. In deriving the updated equations, $L_j$ is treated as the estimation of $D_j$, so the updated Eq. (4) is replaced with the following equations:

$$\lambda_j = \text{tr} \left( \sum_{i=4}^n A^{-1} \Omega_j \right) / d \sum_{i=1}^n z_i^m$$  \hspace{1cm} (8)

$$\Sigma_j = \lambda_j L_j A L_j^T, \quad j = 1, \ldots, c$$  \hspace{1cm} (9)

Case 5: When the parameters of the covariance matrix are the same as for case 4, but the size to be changed, more general elliptical shapes are produced. The covariance matrix is changed to $\Sigma_j = \lambda_j D_j A^T D_j$. In this case, the updated Eq. (4) is replaced with the following equations:

$$\lambda_j = \text{tr} \left( \sum_{i=4}^n A^{-1} \Omega_j \right) / d \sum_{i=1}^n z_i^m$$  \hspace{1cm} (10)

$$\Sigma_j = \lambda_j L_j A L_j^T, \quad j = 1, \ldots, c$$  \hspace{1cm} (11)

Thus, we propose the robust FCML clustering algorithm as follows:

**Robust FCML clustering algorithm**

Step 1: Fix $m \in (1, \infty)$, fix $2 \leq c \leq n$ and fix any $\epsilon > 0$.

- Give an initial partition $z^{(0)}$ and let $s = 1$.
- Estimate the factor $r$ of the penalty term using Eq. (5).

Step 2: Compute $\alpha^{(s)}$ and $u^{(s)}$ with $z^{(s-1)}$ by (1) and (2).

Step 3: Compute $\Sigma^{(s)}$ with $z^{(s-1)}$ and $u^{(s)}$ by (4) or (6) or (7) or (8) and (9) or (10) and (11).

Step 4: Update to $z^{(s)}$ with $\alpha^{(s)}, u^{(s)}$ and $\Sigma^{(s)}$ by (3).

Step 5: Compare $z^{(s)}$ to $z^{(s-1)}$ in a convenient matrix norm.

If $\|z^{(s)} - u^{(s-1)}\| < \epsilon$, Stop.

Else $s = s + 1$, return to Step 2.

The use of these five cases for the robust FCML clustering algorithm is important. The following notations are used to indicate the presentation of different cases, as follows: vvv for case 1, ei for case 2, vi for case 3, eve for case 4 and vve for case 5.
After using the different models: vvv, ei, vi, eve and vve, for the robust FCML clustering algorithm with different cluster numbers, the best model with the optimal number of clusters must be selected. The Bayesian information criterion (BIC) is a very suitable tool for model selection, so it is used for the model selection criterion.

The BIC is a model selection method that is based on the Bayesian method, using the Bayes factor and posterior model probabilities. The main concept is that if there are $M_1, \ldots, M_k$ models, with prior probabilities $p(M_k)$, $k = 1, \ldots, K$, then, by the Bayes theorem, $p(M_k | X) \propto p(X | M_k) p(M_k)$, where $p(X | M_k) = \int p(X | \theta_k, M_k) p(\theta_k | M_k) d\theta_k$ and $p(\theta_k | M_k)$ is the prior distribution of $\theta_k$ and $p(X | M_k)$ is known as the integrated likelihood of the model, $M_k$. The Bayesian approach to model selection is to choose the model that is most likely to be a posterior. To compare two models, $M_1$ and $M_2$, the Bayes factor is defined as the ratio of the two integrated likelihoods, $B_{12} = p(X | M_1) / p(X | M_2)$. However, the main difficulty in using the Bayes factor is the evaluation of the integral that defines the integrated likelihood. For regular models, the integrated likelihood can be approximated by the BIC, using $2 \log p(X | M_k) = 2 \log p(X | \hat{\theta_k}, M_k) - v_k \log(n) = BIC_k$, where $v_k$ is the number of independent parameters to be estimated in $M_k$. A large BIC score indicates strong evidence for the corresponding model, so the BIC score can be used to compare models with different covariance matrix parameterizations and different numbers of clusters for the robust FCML clustering method. All BIC values are calculated from the cluster number, 2, to the given maximum number of clusters for these five different models. Using the plot of the BIC values, the best model with the optimal number of clusters is easily chosen, so a robust FCML clustering framework is constructed, as shown in Fig. 1.

![Fig. 1. Robust FCML clustering framework.](image-url)
4. Numerical Examples and Real Data Applications

In this section, we present several examples that use simulated data. The proposed method is applied to these simulated data sets of different sizes, shapes and orientations. One of the five models: vvv, ei, vi, eve and vve, is chosen as the best one with the optimal number of clusters, using the proposed clustering framework. Yang proposed the FCML, in which the penalized fuzzy c-means (PFCM) is derived using the FCML, so that the well-known fuzzy c-means (FCM) become a special case of the PFCM. However, the FCML with PFCM needs an a priori number of clusters to be assigned and parameter initialization for the factor, $r$, of the penalty term. In 1978, Gustafson and Kessel (GK) studied the effect of different cluster shapes, with the exclusion of the spherical shape, by replacing the Euclidean distance, $d(x_i, u_j) = \|x_i - u_j\|$, in the FCM objective function with the Mahalanobis distance, $d(x_i, u_j) = (x_i - u_j)^T A_j (x_i - u_j)$, where $A_j$ is a positive definite $d \times d$ matrix with a determinate, $\det(A_j) = \rho_j$, that is a fixed constant. This is termed the GK algorithm. Some comparisons of the proposed method with the FCM, PFCM and GK algorithms are given, using the artificial data sets. For real data sets, the method uses iris data, diabetes data, and Wisconsin breast cancer datasets.

Example 1. In the first example, a three-cluster data set with the same shape, but different sample sizes is considered. The data set has 300 data points; 200 data points in one cluster and 50 data points in two other clusters, as shown in Fig. 2. The proposed method is used for the data set and it is found that the BIC values select model ei as the best one, with an optimal cluster number of 3, as shown in Fig. 2. The clustering results for the five models: vvv, ei, vi, eve and vve, are also shown in Fig. 2. The FCM and PFCM are also used in this example. The FCM, PFCM and GK algorithms require the assignment of a cluster number. In fact, these algorithms were implemented using different cluster numbers, but there are always too many misclassifications for these different cluster numbers, except for the cluster number, 3, so only those clustering results for the FCM, PFCM and GK with the cluster number, 3, were used for the comparison, as shown in Fig. 3. It is seen that the results for the PFCM have less misclassifications than those for the FCM, but the proposed method is the best of them. Since the factor, $r$, of the penalty term is effective in the PFCM, the PFCM was also implemented for several values of $r$. The results are shown in Table 1.

![Fig. 2. Data set and different clustering results and BIC values for Example 1.](image-url)
The next examples use data with variation in shape. The proposed method produces good clustering results because the estimation of the factor, $r$, that is suggested is proper and the BIC criterion is used to choose a suitable model.

**Example 2.** This example uses two data sets with line-shaped clusters. In the first data set, the three line-shaped clusters are overlapped in the middle, as shown in the first graph (original) on the left side of Fig. 4. The proposed method is used first for this data set and it is found that the models, $e_i$ and $v_i$, perform badly, but that the other models produce good clustering results with an optimal cluster number of 3, as shown on the left side of Fig. 4. Using the BIC values shown on the right side of Fig. 4, the model, $e_v$, is
best, with an optimal cluster number of 3. Figure 5 shows the clustering results for the FCM, PFCM and GK, and the proposed method is obviously better than the others. In the second data set, the three line-shaped clusters constitute an A-like character, as shown in Fig. 6. For this data set, the models, ei and vi, perform badly, but the other models produce correct cluster numbers and good clustering results. The model, vve, is selected as the best one, with an optimal cluster number of 3. The proposed method is also better than the FCM and PFCM, but GK performs well. The error rates and clustering results are shown in Fig. 7.

Fig. 4. Data set and different clustering results and BIC values for the first data set.

Fig. 5. Error rates and clustering results of our method, FCM and PFCM for the first data set.
Example 3. In this example, a data set is generated from a mixture of three bivariate normal distributions, using an unconstrained covariance matrix. The data set has 200 data points in a spherically-shaped cluster and 50 data points each in two other clusters, as shown in Fig. 8. Using the BIC values from the proposed method, as shown in Fig. 8, the model, vvv, is obviously selected as the best one, with an optimal cluster number of 3. For this data set, it is found that the model, vvv, performs very well, but some other models do not. The error rates and the clustering results for the proposed method, FCM, PFCM and GK are shown in Fig. 9. It is seen that the proposed method still performs well for this data set.
These examples show that the proposed method actually works well for these simulated data sets, with different sizes, shapes and orientations. The next examples use real data sets: iris data, diabetes data and Wisconsin breast cancer datasets.

**Example 4. (Iris data)** The Iris data is a typical test data set for many classification techniques in pattern recognition. The data consists of 50 samples from each of three species of iris flowers (iris setosa, iris virginica and iris versicolor). The attributes of the iris data are the length and the width of the sepal and the petal. The scatterplot for the Iris
data is shown in Fig. 10 (see Wikipedia). Figure 10 shows that there are two clusters in each plot, the larger cluster containing two overlapped clusters. The proposed method was implemented for the Iris data set. By using the random membership initial, $z^{(0)}$, in performing the proposed method, it is found that there are two groups of BIC values for the Iris data set, as shown in Fig. 11. Both groups of BIC values always choose the model, vve, as the best one. It is found that one produces an optimal cluster number of 2, as shown on the left of Fig. 11, but the other produces an optimal cluster number of 3, as shown on the right of Fig. 11. The proposed method was run with 200 times with the random membership initial, $z^{(0)}$, and approximately 70% of the runs produce a group of BIC values with an optimal cluster number of 2 and that approximately 30% of the runs produce a group of BIC values with an optimal cluster number of 3. For the case with an optimal cluster number of 3, the best model, vve, can be more than 85% accurate in correctly detecting the three clusters. In contrast, the second better model, eve, can be more than 82% accurate in correctly detecting the three clusters.

![Fig. 10. Iris data set.](image1)

![Fig. 11. Two groups of BIC values for Iris data set.](image2)
Example 5. (Diabetes dataset) This example uses the diabetes data set, which has 145 observations and 3 attributes (Reaven and Miller). The meaning of the attributes is described as follows: Glucose is the plasma glucose response to oral glucose, Insulin is the plasma insulin response to oral glucose and Sspg is the degree of insulin resistance. This data set is clinically clustered into 3 groups (chemical diabetes, overt diabetes, and normal). Figure 12 is the 3D plot of the diabetes data. It is seen that these clusters are overlapping and are clearly not spherical in shape. The triangular symbol in Fig. 12 denotes the cluster of normal patients. The structure of this cluster has a “fat middle”. The circular and square symbols in Fig. 12 denote the clusters for chemical and overt diabetes, respectively. The shapes of these two clusters resemble two wings. The varying structure of this data means that many clustering methods are unsuitable. The data set was used with the proposed method and the results were compared with those for the FCM and PFCM, for a given cluster number of 3. Firstly, the results for BIC values using the proposed method are shown in Fig. 13. It is seen that the best model is vvv, with an optimal cluster number of 3. This is the cluster number used for the FCM and PFCM. The error rates and clustering results for the proposed method, the FCM and the PFCM are shown in Fig. 14. It is seen that the proposed method performs better than the FCM and the PFCM for the Diabetes dataset.
Example 6. (Wisconsin breast cancer dataset) This example uses the proposed method to analyze the real Wisconsin breast cancer dataset, which is taken from the UCI Machine Learning Repository (also see Wolberg and Mangasarian). There are 569 instances and 32 attributes in this two-cluster data set. The first attribute of this data is the ID number and the second attribute is the diagnosis, which is treated as the true classification (M = malignant, B = benign). Other attributes indicate the real-value features which are computed for each cell nucleus: radius, texture, perimeter, area, etc. The first two indicator attributes were initially ignored and the description of the results using this data from the UCI website was used. The attributes: worst area, worst smoothness and mean texture, were retained and the three-dimensional data was used again, using the proposed method. The proper cluster model, \( v vv \), has an optimal cluster number of 2, as seen in Fig. 15. Using the cluster number, 2, the FCM and PFCM were also implemented using this data. It is seen that the accuracy of the proposed method is 94.5%, which is better than that for the other clustering methods, as shown in Table 2.

<table>
<thead>
<tr>
<th>Clustering method</th>
<th>Our method</th>
<th>PFCM with ( r = 1 )</th>
<th>FCM (PFCM with ( r = 0 ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal accuracy</td>
<td>94.5%</td>
<td>92.1%</td>
<td>91.9%</td>
</tr>
</tbody>
</table>

Fig. 14. Error rates and clustering results from our method, FCM and PFCM.

Fig. 15. BIC values of our method for Wisconsin breast cancer dataset.
Example 7. This example uses a three-component bivariate Gaussian mixture distribution with the parameters (see Chatzis and Varvarigou):

$$\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}; u_1 = (0,3)^T, u_2 = (3,0)^T, u_3 = (-3,0)^T;$$

$$\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}.$$ 

A data set is generated from this Gaussian mixture model with a sample size, $n = 2400$, as shown in Fig. 16. The proposed method is firstly implemented using the data set and the BIC values select the model, vvv, as the best one, with an optimal cluster number of 3, as shown in Fig. 17. In order to determine the effect of noise on the proposed method, 600 noisy points that are drawn from a uniform distribution over the range $[-20,20]$ are added, as shown in Fig. 18. The proposed method is then implemented using this noisy data set and the BIC values also select the model, vvv, as the best one, but with an optimal cluster number of 4, as shown in Fig. 19. For this noisy data set, there is not good clustering. That is, the proposed method is heavily affected by noisy points, for this noisy data set. On adding 600 noisy points that are drawn from a uniform distribution over the range $[-10,10]$, as shown in Fig. 20, the proposed method was also implemented for this noisy data set and the BIC values also select the model, vvv, as the best one, with an optimal cluster number of 4, as shown in Fig. 21. However, for this noisy data set, there is still good clustering, not affected by the noisy points too much. Overall, a comparison with Chatzis and Varvarigou\cite{41} shows that the proposed method is not significantly tolerant of noise or outliers; it may consider all noisy points as another cluster, as shown in Fig. 21.

It is worthy of note that the proposed robust fuzzy CML clustering framework exhibits clustering characteristics that are robustness to parameter initialization, cluster number and different volumes of clusters, but not to noise and outliers. In the literature, there are some studies of clustering characteristics with robustness to noise and outliers that use multivariate $t$-distributions, such as those of Chatzis and Varvarigou\cite{41,42}, Peel and McLachlan\cite{43} and Greselin and Ingrassia.\cite{44} Further research will extend this robust fuzzy CML clustering framework to multivariate $t$-distributions.

Fig. 16. Three-component bivariate Gaussian mixture data set.
A Robust Fuzzy Classification Maximum Likelihood Clustering Framework

Fig. 17. Data set without noises and different clustering results and BIC values for Example 7.

Fig. 18. Gaussian mixture data set with 600 noisy points from a uniform distribution on $[-20,20]$.

Fig. 19. Data set with 600 noisy points and different clustering results and BIC values for Example 7.
5. Conclusions

Using the fuzzy classification maximum likelihood (FCML) with five eigenvalue decomposition models and BIC model selection, we propose a robust FCML clustering framework. The proposed clustering method achieves robust clustering results, with robustness to different cluster volumes and also cluster number and with no parameter initialization problem. Several simulated data and real data sets are used to demonstrate the effectiveness and usefulness of the proposed model. A comparison of the proposed method with the FCM, PFCM, and GK algorithms shows that the proposed method produces better results for the simulated and real data sets. In fact, the FCM algorithm can be derived through the FCML, which means that the proposed model can be seen as a probability model for the FCM. Further study of other relationships and properties, such as stochastic convergence and limiting theorems for the FCM algorithm, are worthwhile, using the proposed model. New fuzzy clustering algorithms may also be developed using the proposed robust FCML clustering framework.
A Robust Fuzzy Classification Maximum Likelihood Clustering Framework

Acknowledgements
The authors are grateful to the anonymous referees for their constructive comments and suggestions to improve the presentation of the paper. This work was supported in part by the National Science Council of Taiwan, under Grant NSC101-2118-M-033-002-MY2.

References