Robust Fuzzy Classification Maximum Likelihood Clustering with Multivariate t-Distributions

Miin-Shen Yang, Yi-Cheng Tian, and Chih-Ying Lin

Abstract

Mixtures of distributions have been used as probability models for clustering data. Classification maximum likelihood (CML) procedure is a popular mixture of maximum likelihood approach to clustering. Yang (1993) extended CML to fuzzy CML (FCML) for a normal mixture model, called FCML-N. However, normal distributions are not robust for outliers. In general, t-distributions should be more robust to outliers than normal distributions. In this paper we consider FCML with multivariate t-distributions and then create a robust clustering algorithm, called FCML-T. To compare with the expectation & maximization for multivariate t-distributions (EM-T), the proposed FCML-T uses a much simpler equation for solving the degrees of freedom. Some numerical and real experimental examples are used to compare the FCML-T with FCML-N, EM for normal mixtures (EM-N) and EM-T. The results demonstrate the superiority and usefulness of the proposed FCML-T algorithm.

Keywords: Mixtures of distributions, fuzzy clustering, classification maximum likelihood (CML), fuzzy CML, multivariate t-distribution, outlier, robustness.

1. Introduction

Cluster analysis is a method for finding clusters of a data set with the most similarity in the same cluster and the largest dissimilarity between different clusters. Cluster analysis has generally become a branch of statistical multivariate analysis [1-2]. According to statistical viewpoint, clustering methods could be generally divided as a probability model based approach and a non-parametric approach. A model-based clustering approach is assumed that a data set follows a mixture model of probability distributions so that a mixture likelihood approach to clustering is used [3-4]. For a nonparametric approach, clustering methods are generally based on an objective function of similarity or dissimilarity measures in which partitioning methods are popularly used [5-6]. The most popular partitioning methods with cluster prototypes are k-means [7-8], and fuzzy c-means (FCM) [9-11].

There are two mixture likelihood approaches to clustering in the literature. One is based on the expectation & maximization (EM) algorithm [12] where the EM algorithm for normal mixture models (EM-N) is the most popular [13]. However, it is not robust for outliers. Peel and MacLaren [14] first considered a robust EM based on t-distribution mixture models (EM-T), and Liu and Rubin [15] and Shoham [16] made more studies. Recently, Lo and Gottardo [17] made advanced study on this robust EM-T. Another one is based on the classification maximum likelihood (CML) [18-20]. Scott and Symons [18] first proposed the CML procedure. Afterwards, Yang [21] extended the original CML to the so-called fuzzy CML (FCM) based on fuzzy c-partitions for normal mixture models, called FCML-N. In the literature, we find that there are several applications of the FCML-N [22-26]. Recently, Yang et al. [27] extended the FCML-N with the eigenvalue decomposition of a covariance matrix for different cluster shapes and also Bayesian information criterion (BIC) for a selection of cluster number. However, these FCML-Ns are not enough as a robust clustering for outliers.

In this paper we construct a robust clustering based on FCML with multivariate t-distributions, called FCML-T. There are the following three main contributions from the proposed FCML-T: (i) It gives much more robustness than FCML-N for outliers; (ii) It provides an estimate for the factor of the penalty term in the FCML-T objective function; (iii) It considers a much simpler equation for solving the degrees of freedom than EM-T. Of course, the FCML-T needs to estimate more parameters so that it spends more running time than FCML-N, but we can get much higher accuracy than FCML-N. The remainder of the paper is organized as follows. In Section 2, we review the FCML-N proposed by Yang [21]. In Section 3, we construct the robust...
FCML-T algorithm. In Section 4, we give some numerical and real comparisons of the proposed FCML-T with FCML-N, EM-N and EM-T. Finally, conclusions are stated in Section 5.

2. Fuzzy Classification Maximum Likelihood Clustering

The classification maximum likelihood (CML) [18-20] is a mixture maximum likelihood approach to clustering. Let the data set \( X = \{x_1, \ldots, x_n\} \) in \( d \)-dimensional space be a random sample of size \( n \) from a mixture \( f(x; \theta) \) of probability distributions. A partition \( P = (P_1, \ldots, P_c) \) of \( X \) into \( c \) clusters can be represented using mutually disjointed sets \( P_1, \ldots, P_c \) such that \( X = P_1 \cup P_2 \cup \ldots \cup P_c \). Then, the joint probability distribution of \( X \) based on the partition \( P = (P_1, \ldots, P_c) \) can be written as \( \prod_{i=1}^{c} \prod_{j \in P_i} f_i(x_j; \theta) \). A so-called CML approach becomes a maximization problem by choosing \( P = (P_1, \ldots, P_c) \) and \( \theta = (\theta_1, \ldots, \theta_c) \) to maximize the log likelihood function \( \sum_{i=1}^{c} \sum_{j \in P_i} \ln f_i(x_j; \theta) \).

Thus, the CML objective function \( L_{\text{CML}} \) for the data set \( X \) becomes

\[
L_{\text{CML}}(P, \theta) = \sum_{i=1}^{c} \sum_{j \in P_i} \ln f_i(x_j; \theta) .
\]

Since Zadeh [28] proposed fuzzy sets that produced the idea of partial membership of belonging described by a membership function, fuzzy clustering has been widely studied and applied in a variety of substantive areas in which the fuzzy c-mean (FCM) clustering algorithm is the best-known method (see [9-11], [29-32]). In applications of fuzzy clustering, Yang [21] first made the fuzzy extension of the original CML to the fuzzy CML (FCM). Next, we first introduce the concept of hard and fuzzy c-partitions and then the FCML for Gaussian mixture models, i.e. FCML-N.

We know that the CML objective function is \( L_{\text{CML}}(P, \theta) = \sum_{i=1}^{c} \sum_{j \in P_i} \ln f_i(x_j; \theta) \). We now use \( \alpha_1 > 0 \) to denote mixing proportions with the constraint \( \sum_{i=1}^{c} \alpha_i = 1 \) and also use the indicator functions \( z_i(x) = 1 \) if \( x \in P_i \) and \( z_i(x) = 0 \) if \( x \notin P_i \) for all \( x \) in \( X \) and \( i = 1, \ldots, c \). This is generally known as clustering \( X \) into \( c \) clusters using \( z = (z_1, \ldots, z_c) \), and the \( z = (z_1, \ldots, z_c) \) is called the hard c-partition of \( X \). Thus, the CML objective function can be extended to be

\[
L_{\text{CML}}(P, \alpha, \theta) = \sum_{i=1}^{c} \sum_{j \in P_i} \ln \alpha_i f_i(x_j; \theta) = \sum_{i=1}^{c} \sum_{j \in P_i} z_i(x_j) \ln \alpha_i f_i(x_j; \theta)
\]

Consider the extension to allow the indicator functions \( z_i(x) \) to be functions assuming values in the interval \([0, 1]\) such that \( \sum_{i=1}^{c} z_i(x) = 1 \) for all \( x \in X \). In this case, \( z = (z_1, \ldots, z_c) \) is called a fuzzy c-partition of the data set \( X \). Based on this extension and the concept of fuzzy c-partition, Yang [21] proposed the FCML objective function as follows:

\[
J_{\text{FCML}}(z, \alpha, \theta) = \sum_{i=1}^{c} \sum_{j \in P_i} z_{ij} \ln f_i(x_j; \theta) + r \sum_{i=1}^{c} \sum_{j=1}^{m} z_{ij} \ln \alpha_i
\]

subject to \( \sum_{i=1}^{c} \alpha_i = 1 \), \( \alpha_i > 0 \), \( \sum_{j=1}^{m} z_{ij} = 1 \), \( z_{ij} \in [0,1] \), and \( r > 0 \). The \( m > 1 \) represents the degree of fuzziness. Thus, the optimization for \( J_{\text{FCML}}(z, \alpha, \theta) \) is achieved by choosing a fuzzy c-partition \( z \) and a mixing proportion \( \alpha \) and an estimate \( \theta \) to maximize \( J_{\text{FCML}}(z, \alpha, \theta) \). Especially, we can derive the necessary conditions for a maximizer \((z, \alpha, \theta)\) of \( J_{\text{FCML}}(z, \alpha, \theta) \) as follows:

\[
\alpha_i = \frac{\sum_{j=1}^{m} z_{ij}^m}{\sum_{j=1}^{m} \sum_{i=1}^{c} z_{ij}^m}
\]

\[
z_{ij} = z_i(x_j) = \frac{\left( \ln f_i(x_j; \theta) + r \ln(\alpha_i) \right)^{1-r}}{\sum_{i=1}^{c} \left( \ln f_i(x_j; \theta) + r \ln(\alpha_i) \right)^{1-r}}
\]

If we consider \( f_i(x_j; \theta) = f_i(x_j; \mu_i, \Sigma_i) \) as a d-variate normal distribution with mean vector \( \mu_i \) and covariance matrix \( \Sigma_i \). Then the FCML objective function becomes

\[
J^N_{\text{FCML}}(z, \alpha, \mu, \Sigma) = \sum_{i=1}^{c} \sum_{j=1}^{m} z_{ij}^m (-1/2)(d \ln(2\pi) + \ln |\Sigma_i|) + \sum_{i=1}^{c} \sum_{j=1}^{m} z_{ij}^m (-1/2)(x_j - \mu_i)' \Sigma_i^{-1}(x_j - \mu_i) + r \sum_{i=1}^{c} \sum_{j=1}^{m} z_{ij}^m \ln \alpha_i
\]

By Lagrange multiplier, the necessary conditions of \((\mu, \Sigma)\) for the maximum of \( J^N_{\text{FCML}}(z, \alpha, \mu, \Sigma) \) are the following updated equations:

\[
\mu_i = \frac{\sum_{j=1}^{m} z_{ij}^m x_j}{\sum_{j=1}^{m} z_{ij}^m}
\]

\[
\Sigma_i = \frac{\sum_{j=1}^{m} z_{ij}^m (x_j - \mu_i)(x_j - \mu_i)'}{\sum_{j=1}^{m} z_{ij}^m}
\]

Yang [21] had considered the clustering algorithm based on the FCML objective function \( J^N_{\text{FCML}}(z, \alpha, \mu, \Sigma) \), and called FCML-N. Thus, the FCML-N can be summarized as follows:
FCML-N clustering algorithm

Step 1: Fix \( m \in (1, \infty) \), \( 2 \leq c \leq n \) and \( \varepsilon > 0 \). Give a value of \( r \), and an initial partition \( z^{(0)} \) and let \( s = 1 \).

Step 2: Compute \( \alpha^{(s)} \) and \( \mu^{(s)} \) with \( z^{(s-1)} \) by (1) and (3).

Step 3: Compute \( \Sigma^{(s)} \) with \( z^{(s-1)} \) and \( \alpha^{(s)} \) by (4).

Step 4: Update to \( z^{(s)} \) with \( \alpha^{(s)}, \mu^{(s)} \) and \( \Sigma^{(s)} \) by (2).

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

If \( \left\| z^{(s)} - z^{(s-1)} \right\| < \varepsilon \), Stop.

Else \( s = s + 1 \), return to Step 2.

Note that Yang [21] had also considered the clustering algorithm based on the FCML-N objective function \( J_{FCML-N}^N(z, \alpha, \mu, \Sigma) \) with \( \Sigma = I \) and a given value of \( r \) where Yang [21] called the derived algorithm a penalized fuzzy c-means clustering. Although there are some applications of the fuzzy CML-N (see [22-27]), it is not a robust clustering for outliers. In next section, we construct a robust clustering for the FCML based on multivariate t-distributions.

3. The Proposed Robust FCML Clustering with Multivariate t-Distribution

Let the d-dimensional data set \( X = \{x_1, \ldots, x_n\} \) be a random sample of size \( n \) from the multivariate t-distributions \( f(x; \mu, \Sigma, \nu) \). Thus, we have the density function as

\[
f(x; \mu, \Sigma, \nu) = \frac{\Gamma \left( \frac{\nu + d}{2} \right) |\Sigma|^{\frac{d}{2}}}{(\pi\nu)^{\frac{d}{2}}} \left( 1 + \frac{\delta(x; \mu, \Sigma)}{\nu} \right)^{-\frac{\nu + d}{2}}
\]

where \( \delta(x; \mu, \Sigma) = (x - \mu)'\Sigma^{-1}(x - \mu) \) is the Mahalanobis squared distance between \( x \) and \( \mu \) with \( \Sigma \) as the covariance matrix, and \( \Gamma(\nu) \) is the Gamma function with \( \Gamma(\nu) = \int_0^\infty s^{\nu-1} e^{-s} ds \). When \( \nu \) tends to infinity, \( X \) becomes marginally multivariate normal with mean \( \mu \) and covariance matrix \( \Sigma \). The family of t-distributions provides a heavy-tailed alternative to the normal family with mean \( \mu \) and covariance matrix that is equal to a scalar multiple of \( \Sigma \), see for example Fig. 1. Thus, the mixture of multivariate t-distributions with \( c \)-components has the density function as

\[
f(x; \psi) = \sum_{i=1}^c \alpha_i f(x; \mu_i, \Sigma_i, \nu_i)
\]

where \( \psi = (\alpha_1, \ldots, \alpha_c, \theta_0' , \nu' , \theta_1, \ldots, \theta_c)' \), \( \nu = (\nu_1, \ldots, \nu_c)' \), and \( \theta_0 = (\theta'_1, \ldots, \theta'_c)' \), and \( \theta_i \) contains the elements of \( \mu_i \) and \( \Sigma_i \), \( i = 1, \ldots, c \).

It is well known that the mixture of multivariate t-distributions can be considered as a scale mixture of normal distributions. Let \( Y \) be the latent variable such that

\[
X | y \sim N(\mu, \Sigma / y) \quad Y \sim G(\nu, \nu / 2)
\]

where \( N(\mu, \Sigma / y) \) denotes the normal distribution with the density function

\[
\phi(x; \mu, \Sigma / y) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left( \frac{1}{2} (x - \mu)'(\Sigma / y)^{-1}(x - \mu) \right)
\]

where \( G(\nu, \nu / 2) \) denotes the gamma distribution with the density function

\[
f(y; \nu, \nu / 2) = \frac{\nu^\nu y^{\nu-1}}{\Gamma(\nu)} e^{-\nu y}, y > 0.
\]

Then, marginally \( X \) has a t-distribution with its density function as

\[
f(x; \mu, \Sigma, \nu) = \frac{\Gamma \left( \frac{\nu + d}{2} \right) |\Sigma|^{\frac{d}{2}}}{(\pi\nu)^{\frac{d}{2}}} \left( 1 + \frac{\delta(x; \mu, \Sigma)}{\nu} \right)^{-\frac{\nu + d}{2}}
\]

Let \( \mathbf{U} = (y_{ij})_{n \times c} \). Thus, the log likelihood of FCML with the mixture of multivariate t-distributions, i.e. FCML-T, can be written as:

\[
J_{FCML}^T(z, \theta, \nu, \mathbf{U}) = \sum_{i=1}^n \sum_{j=1}^c z_{ij} \ln \phi(x_i; \mu_{i}, \Sigma_{ij} / y_{ij}) f(y_{ij}; \nu, \nu / 2) + r \sum_{i=1}^n \sum_{j=1}^c z_{ij} \ln \alpha_i
\]

where \( r \geq 0 \), and it becomes the following form:

\[
J_{FCML}^T(z, \theta, \nu, \mathbf{U}) = \sum_{i=1}^n \sum_{j=1}^c z_{ij} \left( \frac{\nu}{2} \ln \left( \frac{\nu}{2} \right) + \frac{\nu - 1}{2} \ln y_{ij} - \ln \left( \frac{\Gamma \left( \frac{\nu}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right)} \right) - \frac{\nu}{2} \ln (2\pi) - \ln y_{ij} \right)
\]

\[\frac{1}{2} \ln |\Sigma_{ij}| \left( (x_i - \mu_{ij})'\Sigma_{ij}^{-1}(x_i - \mu_{ij}) + v_i \right) + r \sum_{i=1}^n \sum_{j=1}^c z_{ij} \ln \alpha_i
\]

By Lagrange multiplier, the necessary conditions of
(z, θ, v, U) to maximize \( J^F_{FCLM} (z, \theta, v, U) \) are as follows:

\[
y_{ij} = (v_i + d - 2) / (\{x_j - \mu_j\}^T \Sigma^{-1} (x_j - \mu_j) + v_i) \tag{5}\]

\[
\mu_i = \sum_{j=1}^{n} z_{ij} y_{ij} / \sum_{j=1}^{n} z_{ij}^r y_{ij} \tag{6}\]

\[
\Sigma_i = \sum_{j=1}^{n} z_{ij}^r (x_j - \mu_j)(x_j - \mu_j)^T / \sum_{j=1}^{n} z_{ij}^r \tag{7}\]

By following the way from Kent et al. [33], we may replace the divisor \( \sum_{j=1}^{n} z_{ij}^r \) in (7) with \( \sum_{j=1}^{n} z_{ij}^r y_{ij} \). Thus, the update equation for covariance matrix becomes

\[
\Sigma_i = \sum_{j=1}^{n} z_{ij}^r (x_j - \mu_j)(x_j - \mu_j)^T / \sum_{j=1}^{n} z_{ij}^r y_{ij} \tag{7'}\]

In addition, the estimates of the degrees of freedom \( v_i \) can be derived as follows:

\[
\frac{\partial J^F_{FCLM} (z, \theta, v, U)}{\partial v_i} = \sum_{j=1}^{n} z_{ij}^r \left[ \ln \left( \frac{v_i}{2} \right) + \frac{1}{2} \ln \left( \frac{v_i}{2} \right)^T - \frac{1}{2} \psi \left( \frac{v_i}{2} \right) \right] = 0 \tag{8}\]

where \( \psi(s) \) is the digamma function with \( \psi(s) = \partial \ln \Gamma(s) / \partial s \).

We next consider the estimation for the factor \( r \) of the penalty term \( \sum_{i=1}^{n} \sum_{j=1}^{n} z_{ij}^r \ln \alpha_i \) in the objective function \( J^F_{FCLM} (z, \theta, v, U) \) as follows. Since the penalty term is a function of mixing proportions \( \alpha_i \), the factor \( r \) should be proportional to the impact of mixing proportions \( \alpha_i \) on \( J^F_{FCLM} (z, \theta, v, U) \). On the other hand, the true values of mixing proportions \( \alpha_i \) depend on the cluster structure of the data set, the value of \( r \) could be estimated according to its data structure. In this sense, we may use the correlation coefficient between data attributes of the data set to estimate \( r \). Here, we borrow the concept from Yang et al. [27]. Let \( X = \{x_i, \cdots, x_n\} \) be a set of \( n \) data points in a \( d \)-dimensional Euclidean space and let \( y_i = (y_{i1}, \cdots, y_{id}) \) be the data vector of \( i \)th dimension for \( X \). We then calculate the correlation coefficients between these attribute vectors with

\[
\rho_{x_i y_j} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(y_{jk} - \bar{y}_j)}{\sqrt{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^{n} (y_{jk} - \bar{y}_j)^2}}, i \neq j. \tag{9}\]

We start the value of \( r \) with 1/2 and then increase the value with positive correlation coefficients and decrease the value with negative correlation coefficients. Thus, we give the following estimate for \( r \):

\[
r = \frac{1 + \sum_{j=1}^{n} \sum_{k=j+1}^{n} \rho_{x_j y_k}}{2} \tag{9'}\]

Furthermore, we let \( r = 1 \), if \( r \geq 1 \), but we let \( r = 0 \), if \( r \leq 0 \). We now propose the FCML-T clustering algorithm as follows:

**FCML-T clustering algorithm**

Step 1: Fix \( m \in (1, \infty) \), \( 2 \leq c \leq n \) and \( \varepsilon > 0 \). Give an initial partition \( z^{(0)}, v^{(0)}, Y^{(0)} \), and let \( s = 1 \).

Step 2: Estimate the factor \( r \) using (9).

Step 3: Compute \( \alpha^{(s)} \) with \( z^{(s-1)} \) by (1).

Step 4: Compute \( \mu^{(s)} \) with \( z^{(s-1)} \) and \( Y^{(s-1)} \) by (6).

Step 5: Compute \( \Sigma^{(s)} \) with \( z^{(s-1)}, Y^{(s-1)} \) and \( \mu^{(s)} \) by (7').

Step 6: Compute \( Y^{(s)} \) with \( v^{(s-1)}, \mu^{(s)} \) and \( \Sigma^{(s)} \) by (5).

Step 7: Find the solution of \( v^{(s)} \) of (8).

Step 8: Update to \( z^{(s)} \) with \( \alpha^{(s)}, \mu^{(s)}, \Sigma^{(s)} \) and \( v^{(s)} \) by (2).

Step 9: Compare \( z^{(s)} \) to \( z^{(s-1)} \) in a matrix norm.

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

Else \( s = s + 1 \), return to Step 3.

We give an example to demonstrate the effect of our proposed FCML-T.

**Example 1:** In this example, we consider a data set generated from a three-component bivariate normal mixture

\[
f(x; \psi) = \sum_{i=1}^{3} \alpha_i N(u_i, \Sigma_i) \]

with the following parameters:

\[
\alpha_i = \frac{5}{36}, \alpha_2 = \frac{11}{36}, \alpha_3 = \frac{20}{36} \quad \text{and} \quad u_i = (6, 0)^T, \]

\[
u_1 = (-6, 0)^T, v_2 = (0, 9)^T \quad \text{and} \quad \Sigma_1 = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 10 & 0 \\ 0 & 4 \end{pmatrix}. \]

The data set is generated from this mixture model with a sample size \( n=360 \). We add an outlier \((25, 25)\) to this data set as shown in Fig. 2. We perform FCML-T for the data set with various value of \( r \) by considering 50 random initials. The results are shown in Table 1. We find that the FCML-T has the smallest error rate when \( r = r^*=0.7898 \), where \( r^* \) is estimated by equation (10).

### 4. Numerical and Real Data Experiments

In this section, several examples are used to demon-
-strate the superiority and usefulness of the proposed FCML-T clustering algorithm. In all examples, we give $m=1.3$. We also use the estimate of $\rho$ with equation (9) for the FCML-N clustering algorithm in order to give comparisons.

**Example 2** (3-normal-component data set): We first consider a data set generated from a three-component bivariate normal mixture $f(x; \psi) = \sum_{i=1}^{3} \alpha_i N(u_i, \Sigma_i)$ with the following parameters:

$$\alpha_i = \alpha_2 = \alpha_3 = \frac{1}{3}; u_1 = (5, 0)^T, u_2 = (-5, 0)^T, u_3 = (0, 5)^T;$$

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

The data set, as shown in Fig. 3(a), is generated from this normal mixture model with a sample size $n=600$. We perform EM-N, EM-T, FCML-N and FCML-T for the data set by considering 50 random initials. We record the average error rates for the four methods. We find that there are almost the same error rates for the four algorithms where both EM-N and FCML-N present better than both EM-T and FCML-T for this 3-normal-component data set. In considering the robustness for outliers, we add an outlier (20, 20) to this data set, as shown in Fig. 3(b). The results are shown in Fig. 4 and Table 2. Although the error rates of EM-N and FCML-N are less than those of EM-T and FCML-T for the 3-normal-component data set without outlier, the error rates of EM-N and FCML-N rapidly increase to 0.214 and 0.232, respectively, for the data set with the outlier. We find that the error rate of FCML-T is stably with 0.004 which is the smallest among all methods. The FCML-T actually presents more robust than FCML-N, EM-N and EM-T for the data set with the outlier.

**Example 3:** In this example, the numerical data set from Cuesta-Albertos et al. [34] and McLachlan et al. [35] is considered where the data set is generated from a three-component bivariate normal mixture $f(x; \psi) = \sum_{i=1}^{3} \alpha_i N(u_i, \Sigma_i)$ with the following parameters:

$$\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}; u_1 = (-2, 0)^T, u_2 = (0, 0)^T, u_3 = (2, 0)^T;$$

$$\Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 0.2 & 0 \\ 0 & 2 \end{pmatrix}.$$

The data set is generated from this normal mixture model with a sample size $n=600$ and added with 20 noisy points from the uniform distribution on the set $\{(x_1, x_2) \in [-5, 5] \times [-8, 8]: x_1 < -4, or > 4, or x_2 < -5 or > 5\}$ as shown in Fig. 5. We perform FCML-T and FCML-N for the data set by considering 50 random initials. We find that the error rate of FCML-T is 0.045 and the error
rate of FCML-N is 0.359. We also demonstrate the clustering results of FCML-T and FCML-N as shown in Figs. 6 and 7, respectively. We find that FCML-T is much more robust than FCML-N for this uniformly noisy data set (with background outliers). We also compare FCML-T with EM-N and EM-T. The results are shown in Table 3. We find that the EM-T has the smallest error rate among these four algorithms for this data set, but the FCML-T still has very small error rate.

![Data set generated from the three bivariate normal mixture with background outliers.](image1)

Figure 5. Data set generated from the three bivariate normal mixture with background outliers.

![Clustering results from FCML-T.](image2)

Figure 6. Clustering results from FCML-T.

![Clustering results from FCML-N.](image3)

Figure 7. Clustering results from FCML-N.

Table 3 Average error rate of EM-N, EM-T, FCML-N and FCML-T.

<table>
<thead>
<tr>
<th>Data set of Fig. 5</th>
<th>EM-N</th>
<th>EM-T</th>
<th>FCML-N</th>
<th>FCML-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average error rate</td>
<td>0.653</td>
<td>0.031</td>
<td>0.359</td>
<td>0.045</td>
</tr>
</tbody>
</table>

Example 4: In this example, we continue Example 3. We add 20 points from the uniform distribution on the set \( \{(x_1, x_2) \in [7, 8] \times [-8, -7] \} \), as shown in Fig. 8 with local contamination. We perform EM-N, EM-T, FCML-N and FCML-T for the data set by considering 50 random initials. The results are shown in Figs. 9 and 10. We find that the error rate of FCML-T is 0.038, the error rate of FCML-N is 0.322, the error rate of EM-N is 0.365 and the error rate of EM-T is 0.360. Thus, the FCML-T presents much more robust than EM-N, FCML-N and EM-T for this 3-component normal data set with contamination.

![Data set generated from the mixture of three bivariate normal distributions with contamination.](image4)

Figure 8. Data set generated from the mixture of three bivariate normal distributions with contamination.

![Clustering results from FCML-T; (b) Clustering results from EM-T; (c) Clustering results from EM-N.](image5)

Figure 9. (a) Clustering results from FCML-T; (b) Clustering results from FCML-N; (c) Clustering results from EM-T; (d) Clustering results from EM-N.

![Error rates of EM-N, EM-T, FCML-N and FCML-T for 3-normal with contamination.](image6)

Figure 10. Error rates of EM-N, EM-T, FCML-N and FCML-T for 3-normal with contamination.

Example 5: This example uses the data set from Chatzis and Varvarigou [36] with a three-component bivariate...
normal mixture distribution \( f(x; \psi) = \sum_{i=1}^{3} \alpha_i N(u_i, \Sigma_i) \) having the following parameters:
\[
\alpha_1 = \alpha_2 = \alpha_3 = \frac{1}{3}, u_1 = (0, 3)^T, u_2 = (3, 0)^T, u_3 = (-3, 0)^T; \\
\Sigma_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \Sigma_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}
\]

A data set is generated from this normal mixture model with a sample size, \( n=2400 \). We also add 600 noisy points from a uniform distribution over the range \([-20, 20]\) to the data set, as shown in Fig. 11. We perform EM-N, EM-T, FCML-N and FCML-T for the data set with 50 random initials. The average error rates for the four algorithms are shown in Fig. 12. We find that the error rate of FCML-T is 0.010 and the error rate of EM-T is 0.014, but the error rate of FCML-N is 0.547 and the error rate of EM-N is 0.503. We find that FCML-N and EM-N are heavily affected by these 600 noisy points, but FCML-T and EM-T are not affected where FCML-T presents the most robustness.

Example 6: In this example, we consider a data set generated from a mixture of three bivariate t-distributions:
\[
\frac{1}{3} \begin{pmatrix} 5 \\ 0 \end{pmatrix} \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} -5 \\ 0 \end{pmatrix} \begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 0 \\ 5 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}
\]

We generate 600 data points from this t-distribution mixture model, as shown in Fig. 13. We implement EM-N, EM-T, FCML-N and FCML-T for the data set by considering 50 random initials. We record the average error rate of each method. We find that the error rates of EM-N, EM-T, FCML-N and FCML-T are 0.339, 0.027, 0.260 and 0.034, respectively. The clustering results are shown in Fig. 14 and Fig. 17 and Table 4. Thus, EM-T presents the most robust among the four algorithms. But if we consider the robustness for noises by adding 200 noisy points from a uniform distribution over the range \([-20, 20]\) as shown in Fig. 15. We implement EM-N, EM-T, FCML-N and FCML-T for the data set by considering 50 random initials. The clustering results are shown in Fig. 16, Fig. 17 and Table 4. We find that the error rate of EM-N is 0.372, the error rate of FCML-N is 0.653, the error rate of EM-T is 0.332 and the error rate of FCML-T is 0.025. Thus, FCML-T becomes the most robust among the four algorithms for this t-distribution mixture with noises.

Example 7 (Iris data): The Iris data are a typical test data set for most classification techniques in pattern recognition. The data consists of 50 samples from each of three species of iris flowers (iris setosa, iris virignica and iris versicolor) [37]. The attributes of Iris data were measured the length and the width of sepal and petal. The scatterplot for Iris data is shown in Fig. 18(a) (see
Figure 15. Data set generated from the mixture of three bivariate t distributions with 200 noisy points.

Figure 16. (a) Clustering results from FCML-T; (b) Clustering results from FCML-N; (c) Clustering results from EM-T; (d) Clustering results from EM-N.

Figure 17. Error rates of EM-N, EM-T, FCML-N and FCML-T for t-distribution and t-distribution with 200 Noisy points.

Table 4. Average error rates of EM-N, EM-T, FCML-N and FCML-T.

<table>
<thead>
<tr>
<th></th>
<th>EM-N</th>
<th>EM-T</th>
<th>FCML-N</th>
<th>FCML-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-distribution</td>
<td>0.339</td>
<td>0.027</td>
<td>0.260</td>
<td>0.034</td>
</tr>
<tr>
<td>t-distribution with noises</td>
<td>0.372</td>
<td>0.332</td>
<td>0.653</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Wikipedia for Irish data). From Fig. 18(a), we can find that there are two clusters in each plot where the larger cluster contains two overlapped clusters. We perform EM-N, EM-T, FCML-N and FCML-T for the Iris data set. By considering 100 random initials for performing these methods, we record the average error rates of each method. We find that the error rate of FCML-T is 5.21% which is smallest among all methods. We then add an outlier (50, 50, 50, 50) to the Iris data set and it is denoted by Iris*. We perform EM-N, EM-T, FCML-N and FCML-T for this data set by considering 100 random initials. The results are shown in Fig.18 (b) and Table 5. We find that the error rate of FCML-T is 6.50% which is also the smallest among these methods but the error rates of FCML-N, EM-N and EM-T increase to 66.67%, 62.73% and 54.60%, respectively. The FCML-T actually presents much more robust than FCML-N, EM-N and EM-T for the iris data and the iris data with outlier.

Table 5. Average error rates of EM-N, EM-T, FCML-N and FCML-T.

<table>
<thead>
<tr>
<th></th>
<th>EM-N</th>
<th>EM-T</th>
<th>FCML-N</th>
<th>FCML-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris</td>
<td>22.23%</td>
<td>12.52%</td>
<td>24.47%</td>
<td>5.21%</td>
</tr>
<tr>
<td>Iris*</td>
<td>62.73%</td>
<td>54.60%</td>
<td>66.67%</td>
<td>6.50%</td>
</tr>
</tbody>
</table>

Example 8 (Flea beetle data set): In this example, we consider the flea beetle data from Lubischew [38], where “1” denotes the cluster of heikertingeri, “2” denotes the cluster of heptapotamica and “3” denotes the cluster of concinna, as shown in Fig. 19(a). The data set has 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 = 7.5°). We implement the EM-N, EM-T, FCML-N and FCML-T algorithms for this data set with
100 random initials. The error rates are shown in Fig. 19(b) and Table 6. We find that the proposed FCML-T performs the best for this real data set of flea beetle with the error rate only 0.057, where the EM-T also presents very well, but both FCML-N and EM-N present badly.

![Figure 19. (a) Flea beetle data set; (b) Error rates of EM-N, EM-T, FCML-N and FCML-T.](image)

Table 6. Average error rates of EM-N, EM-T, FCML-N and FCML-T.

<table>
<thead>
<tr>
<th>Error rates</th>
<th>EM-N</th>
<th>EM-T</th>
<th>FCML-N</th>
<th>FCML-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flea beetle</td>
<td>0.188</td>
<td>0.069</td>
<td>0.278</td>
<td>0.057</td>
</tr>
</tbody>
</table>

Example 9 (Crab data set): In this example, we consider the crab data set of Campbell and Mahon [39] on the genus Leptograpsus which are consisted of measures over a sample of 100 blue crabs. Each specimen has 5 measurements: the width of the front lip (FL), the rear width (RW), the length along the midline (CL) and the maximum width (CW) of the carapace, and the body depth (BD) in mm. We give the scatter plot of the second and third features of the data set, as shown in Fig. 20. Peel and McLachlan [14] fitted this data set using both normal and t distributions with equal covariance matrices and equal degrees of freedom. We implement FCML-N and FCML-T for this crab data set and add some outliers. This was done by adding various values to the second variate of the 25th point (see Peel and McLachlan [14]). Then we consider 100 random initials for performing FCML-N and FCML-T. We count their error rates and compare with EM-N and EM-T. The results are shown in Table 7. We find that the error rates from FCML-T are always smaller than FCML-N, EM-N and EM-T. As a whole, FCML-T presents much more robustness than FCML-N, EM-N and EM-T.

![Figure 20. Plot of the second and third features for crab data.](image)

Table 7. Average error rates of EM-N, EM-T, FCML-N and FCML-T.

<table>
<thead>
<tr>
<th>Constant</th>
<th>EM-N</th>
<th>EM-T</th>
<th>FCML-N</th>
<th>FCML-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>-15</td>
<td>49%</td>
<td>19%</td>
<td>49.98%</td>
<td>13.09%</td>
</tr>
<tr>
<td>-10</td>
<td>49%</td>
<td>19%</td>
<td>49%</td>
<td>10.32%</td>
</tr>
<tr>
<td>-5</td>
<td>21%</td>
<td>20%</td>
<td>23.56%</td>
<td>12.44%</td>
</tr>
<tr>
<td>0</td>
<td>19%</td>
<td>18%</td>
<td>9.55%</td>
<td>5.42%</td>
</tr>
<tr>
<td>5</td>
<td>21%</td>
<td>20%</td>
<td>15.55%</td>
<td>10.02%</td>
</tr>
<tr>
<td>10</td>
<td>50%</td>
<td>20%</td>
<td>45.93%</td>
<td>13.21%</td>
</tr>
<tr>
<td>15</td>
<td>47%</td>
<td>20%</td>
<td>47.99%</td>
<td>13.56%</td>
</tr>
<tr>
<td>20</td>
<td>49%</td>
<td>20%</td>
<td>50%</td>
<td>14.05%</td>
</tr>
</tbody>
</table>

5. Conclusions

According to the fuzzy classification maximum likelihood (FCML), we proposed an FCML based on multivariate t-distributions and then construct a robust clustering method, called FCML-T. It considered an estimate for the factor of the penalty term in the FCML-T objective function and also had a simpler equation for solving the degrees of freedom than EM-T. To demonstrate the superiority and usefulness of the proposed FCML-T algorithm, we used five simulated data sets and three real data sets for comparing FCML-T with FCML-N, EM-N and EM-T. According to those comparisons, the FCML-T actually presents better results. We find that the FCML-T presents much robust to outliers, background outliers, noisy points and contamination. Although the method spends more running time than FCML-N because of more parameter estimations, we can actually get much higher accuracy than FCML-N.

On the other hand, the Internet has considered a necessity by most people. There are many resources of knowledge and data patterns on the Web. Web mining becomes an important tool for discovering different knowledge and patterns from the Web on the Internet. For example, Web usage mining is used as a process for finding out what users are looking for on the Internet. In general, clustering is considered as an unsupervised learning for Web mining, such as Web documents from World Wide Web and HTML files [40]. Since knowledge and data patterns from the Web are often imprecise, vague and incomplete, fuzzy clustering can be suitably used to handle such Web data [41]. Recently, fuzzy Web mining techniques become more important, and are
widely studied [42-43]. Because our proposed FCML-T is presented as a robust fuzzy clustering, the application of FCML-T to fuzzy Web mining will be our next research topic. Furthermore, since the FCML-T is still necessary to give a priori cluster number, finding a learning schema with automatically obtaining an optimal cluster number for the FCML-T may also be an interesting research topic.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments in improving the presentation of this paper.

References


Miin-Shen Yang received the BS degree in mathematics from the Chung Yuan Christian University, Chungli, Taiwan, in 1977, the MS degree in applied mathematics from the National Chiao-Tung University, Hsinchu, Taiwan, in 1980, and the PhD degree in statistics from the University of South Carolina, Columbia, USA, in 1989.

In 1989, he joined the faculty of the Department of Mathematics in the Chung Yuan Christian University as an Associate Professor, where, since 1994, he has been a Professor. From 1997 to 1998, he was a Visiting Professor with the Department of Industrial Engineering, University of Washington, Seattle, USA. During 2001-2005, he was the Chairman of the Department of Applied Mathematics in the Chung Yuan Christian University and now, he is the Director of Chaplain’s Office. His current research interests include applications of statistics, fuzzy clustering, neural fuzzy systems, pattern recognition and machine learning.

Dr. Yang was an Associate Editor of the IEEE Transactions on Fuzzy Systems (2005-2011), and is an Associate Editor of the Applied Computational Intelligence & Soft Computing and Editor-in-Chief of Advances in Computational Research.

Yi-Cheng Tian received the BS degree in mathematics from the Feng Chia University, Taichung, Taiwan, in 1987, the MS degree in applied mathematics from the National Cheng Chi University, Taipei, Taiwan, in 1989.

Since 2009, he has been a Ph.D. student in the Department of Applied Mathematics at Chung Yuan Christian University, Chung-Li, Taiwan. His research interests include cluster analysis, fuzzy clustering and data analysis.

Chih-Ying Lin received the BS degree in mathematics from the Chung Yuan Christian University, Chung-Li, Taiwan, in 2003, the MS degree in applied mathematics from the National Dong Hwa University, Hualien, Taiwan, in 2005, the Ph.D. in the Department of Applied Mathematics at Chung Yuan Christian University, Chung-Li, Taiwan, in 2013. His research interests include cluster analysis, fuzzy clustering and data analysis.