

Gaussian mixture modeling by exploiting the Mahalanobis distance

Dimitrios Ververidis and Constantine Kotropoulos*, *Senior Member, IEEE*

Abstract

In this paper, the expectation-maximization (EM) algorithm for Gaussian mixture modeling is improved via three statistical tests. The first test is a multivariate normality criterion based on the Mahalanobis distance of a sample measurement vector from a certain Gaussian component center. The first test is used in order to derive a decision whether to split a component into another two or not. The second test is a central tendency criterion based on the observation that multivariate kurtosis becomes large if the component to be split is a mixture of two or more underlying Gaussian sources with common centers. If the common center hypothesis is true, the component is split into two new components and their centers are initialized by the center of the (old) component candidate for splitting. Otherwise, the splitting is accomplished by a discriminant derived by the third test. This test is based on marginal cumulative distribution functions. Experimental results are presented against seven other expectation-maximization variants both on artificially generated data-sets and real ones. The experimental results demonstrate that the proposed EM variant has an increased capability to find the underlying model, while maintaining a low execution time.

Index Terms

Expectation maximization algorithm (EM), Gaussian mixture models (GMM), normality criterion, distribution of Mahalanobis distance, multivariate kurtosis.

I. INTRODUCTION

The Expectation-Maximization algorithm (EM) is widely used to find the parameters of a mixture of Gaussian probability density functions (pdfs) or briefly Gaussian components that fits the sample measurement vectors in maximum likelihood sense [1].

* Corresponding author. D. Ververidis and C. Kotropoulos are with the Dept. of Informatics, Aristotle Univ. of Thessaloniki, Box 451, Thessaloniki 54124, Greece. E-mail:{jimver,costas}@aiaa.csd.auth.gr

However, the EM algorithm is not limited only to find the parameters of a density mixture model. It can be used to 1) detect samples that deviate from a priori known distributions [1], [2]; 2) find the weight parameters in the *re-weight least squares* method [1]; 3) calculate the parameters of Hidden Markov Models (HMMs) with *Baum-Welch* or *forward-backward* algorithm [3]; 4) select features, i.e. to find a feature subset that achieves the lowest prediction error [4].

Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ be the observed data, i.e. \mathcal{X} is a set of random vectors (R.V.s.) \mathbf{x}_i , where \mathbf{x}_i belongs to an arbitrary sample space \mathcal{X} . Let also $g(\mathcal{X}|\Xi)$ be a certain function of \mathcal{X} and parameters Ξ , which is often called as *likelihood function*. The *log-likelihood function*, i.e. the natural logarithm of $g(\mathcal{X} | \Xi)$ is preferred instead of $g(\mathcal{X} | \Xi)$ in order to avoid over- or underflow errors, i.e.

$$\Lambda_1(\mathcal{X}|\Xi) = \ln \underbrace{\prod_{i=1}^N f(\mathbf{x}_i|\Xi)}_{g(\mathcal{X}|\Xi)} = \sum_{i=1}^N \ln f(\mathbf{x}_i | \Xi), \quad (1)$$

where $f(\mathbf{x}|\Xi)$ is the pdf of \mathbf{x} . The target is to find the optimal parameter vector Ξ , denoted as Ξ^* , so that $\Lambda_1(\mathcal{X}|\Xi)$ is maximized. Since a closed solution for Ξ^* can not be found in general, the EM algorithm is used to iteratively find Ξ by applying two steps, the so-called expectation step (E-step) and the maximization step (M-Step).

By introducing *unobserved variables* $h_q(\mathbf{x}_i)$ to denote the probability of a sample measurement vector \mathbf{x}_i belongs to the q th component, $q = 1, 2, \dots, Q$, the conditional expectation of the log-likelihood function is defined as

$$\Lambda_2(\mathcal{X}|\Xi) = \sum_{i=1}^N \sum_{q=1}^Q h_q(\mathbf{x}_i) \ln [\pi_q f_q(\mathbf{x}_i | \Xi_q)], \quad (2)$$

where $f_q(\mathbf{x}_i | \Xi_q)$ is the pdf of the q th component with parameters $\Xi_q \subset \Xi$, and $\pi_q \in [0, 1]$ are the priors of each density function subject to $\sum_{q=1}^Q \pi_q = 1$. The EM algorithm can be considered as a “soft” version of the k -means clustering [5]. In k -means, each sample measurement vector is assigned to a cluster with probability either 0 or 1, whereas in EM, the probability $h_q(\mathbf{x}_i)$ that a sample measurement vector $\mathbf{x}_i \in \mathcal{X}$ belongs to the q th Gaussian component lies in $[0, 1]$. In the special case where each density function $f_q(\mathbf{x}|\Xi)$ is a Gaussian one denoted as

$$p(\mathbf{x} | \boldsymbol{\mu}_q, \mathbf{S}_q) = (2\pi)^{-D/2} (|\mathbf{S}_q|)^{-0.5} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_q)^T \mathbf{S}_q^{-1}(\mathbf{x} - \boldsymbol{\mu}_q)\right\} \quad (3)$$

where $|\cdot|$ is the determinant of the matrix inside the delimiters and D is the dimension cardinality of \mathbf{x} , the pdf of \mathbf{x} is the Gaussian mixture model (GMM)

$$p(\mathbf{x}_i | \Xi) = \sum_{q=1}^Q \pi_q p(\mathbf{x}_i | \boldsymbol{\mu}_q, \mathbf{S}_q). \quad (4)$$

The following parameters should be estimated for each component: the prior π_q^r , the sample mean vector $\boldsymbol{\mu}_q^r$, and the sample dispersion matrix \mathbf{S}_q^r , where r denotes the r th iteration of the EM algorithm. The parameters of the GMM can be collected to a parameter vector $\boldsymbol{\Xi}^r = \{\pi_q^r, \boldsymbol{\mu}_q^r, \mathbf{S}_q^r\}_{q=1}^Q$. The initial parameter vector $\boldsymbol{\Xi}^1$ is randomly chosen or selected by the methods analyzed in Section I-C. Next, the parameters $\pi_q^r, \boldsymbol{\mu}_q^r, \mathbf{S}_q^r$ for $r \geq 2$ are re-estimated using the E- and M-Steps [1]:

E-step: The probability that each vector \mathbf{x}_i belongs to q th component is calculated by

$$h_q^r(\mathbf{x}_i) = \frac{\pi_q^{r-1} p(\mathbf{x}_i | \boldsymbol{\mu}_q^{r-1}, \mathbf{S}_q^{r-1})}{\sum_{q'=1}^Q \pi_{q'}^{r-1} p(\mathbf{x}_i | \boldsymbol{\mu}_{q'}^{r-1}, \mathbf{S}_{q'}^{r-1})}, \quad (5)$$

M-step: The prior, the sample mean vector, and the sample dispersion matrix of each component are recalculated by using $h_q^r(\mathbf{x}_i)$:

$$\pi_q^r = \frac{1}{N} \sum_{i=1}^N h_q^r(\mathbf{x}_i), \quad (6)$$

$$\boldsymbol{\mu}_q^r = \frac{\sum_{i=1}^N h_q^r(\mathbf{x}_i) \mathbf{x}_i}{\sum_{j=1}^N h_q^r(\mathbf{x}_j)}, \quad (7)$$

$$\mathbf{S}_q^r = \frac{\sum_{i=1}^N h_q^r(\mathbf{x}_i) (\mathbf{x}_i - \boldsymbol{\mu}_q^r) (\mathbf{x}_i - \boldsymbol{\mu}_q^r)^T}{\sum_{j=1}^N h_q^r(\mathbf{x}_j)}. \quad (8)$$

The E- and M-steps alternate until the conditional expectation of the log-likelihood function of the GMM defined as

$$\mathcal{L}(\mathcal{X} | \boldsymbol{\Xi}^r) = \sum_{i=1}^N \sum_{q=1}^Q h_q^r(\mathbf{x}_i) \ln \left(\pi_q^r p(\mathbf{x}_i | \boldsymbol{\mu}_q^r, \mathbf{S}_q^r) \right) \quad (9)$$

reaches a local maximum.

For convenience, the EM algorithm for Gaussian mixture modeling is abbreviated as EM algorithm. However, EM is not a panacea, it suffers from two drawbacks: a) the number of Gaussian components Q is usually set a priori, and b) the initialization of the parameters of the Gaussian components $\boldsymbol{\Xi}^1$ affects the final result. Therefore, EM converges to a local optimum of the parameter space. Several techniques have been used in order to escape from local optima. These techniques can be divided into three levels according to the part of the EM algorithm are applied to. These levels are shown in Figure 1. In the 3rd level, techniques for estimating the number of components Q can be found. In the 2nd level, there are techniques that use other EM steps than the standard EM steps in order to escape from local optima.

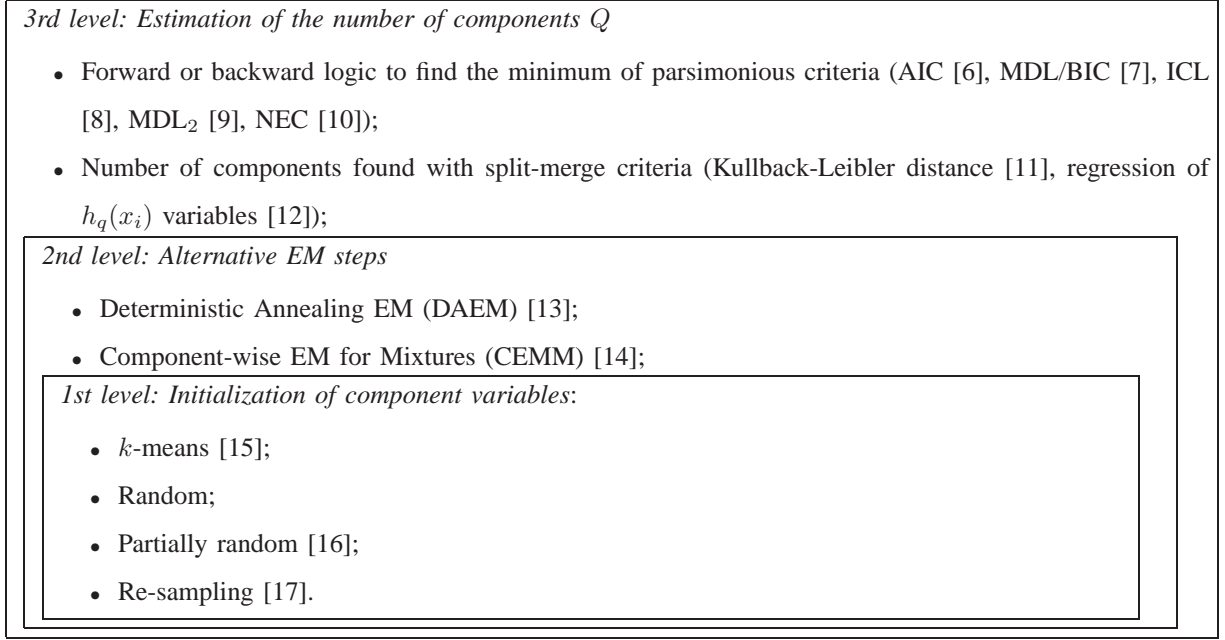


Fig. 1. Techniques to avoid local maxima of the log-likelihood function in EM algorithm.

Finally, in level 1, techniques that initialize the Gaussian component parameters are met. Examples of the techniques from each level will be described next. The examples will be used for the comparison between the state-of-the-art methods against the proposed method in Section V-B.

A. Estimation of the number of components Q (3rd level)

The number of components Q can be found by parsimonious criteria or by split-merge operations applied to components.

Parsimonious criteria relate the log-likelihood function of the model with the number of free parameters in order to prevent an infinite number of mixture components. The forward logic starts from one component in the GMM and increases the number of components by one whenever EM converges, whereas the backward logic starts from many components and removes one component after the convergence of EM [9]. The initial number of components Q in the backward logic can be found as follows. The probability that a component q is not represented in the random initialization is $(1 - \pi_q)^Q$. So, if the desirable probability of successful initialization is at least $1 - \epsilon$, where ϵ is a small positive number, then the initial number of components Q should be [9]:

$$Q = \frac{\log \epsilon}{\log(1 - \pi_{\min})}, \quad \text{where } \pi_{\min} = \min_{q=1}^Q \{\pi_q\}. \quad (10)$$

The drawback in (10) is that π_{\min} should be known a priori. Parsimonious criteria are outlined in Table I, where ν is the dimensionality of Ξ . $\mathcal{L}^*(\mathcal{X} | \Xi)$ is given by (9), whereas for a certain \mathbf{x}_i the greatest $h_q(\mathbf{x}_i)$ attains the value 1, and the remaining $h_q(\mathbf{x}_i)$ tend to zero.

TABLE I
PARSIMONIOUS CRITERIA USED TO PENALIZE THE LOG-LIKELIHOOD FUNCTION

Name	Penalty function to minimize	Reference
Akaike Information criterion (AIC)	$-\mathcal{L}(\mathcal{X} \Xi) + 2\nu$	[6], [18]
Minimum description length or Bayesian Information Criterion (MDL/BIC)	$-\mathcal{L}(\mathcal{X} \Xi) + \frac{\nu}{2} \ln(N)$	[19], [9], [20]
Integrated Completed likelihood	$-\mathcal{L}^*(\mathcal{X} \Xi) + \frac{\nu}{2} \ln(N)$	[19]
Minimum Description Length variant (MDL ₂)	$-\mathcal{L}(\mathcal{X} \Xi) + \frac{Q}{2} \ln \frac{N}{12} + \frac{\nu+Q}{2} + \frac{\nu}{2Q} \sum_{q=1}^Q \ln \frac{N\pi_q}{12}$	[9]
Negative Entropy Criterion (NEC)	$\frac{E(\mathcal{X} \Xi)}{\mathcal{L}(\mathcal{X} \Xi) - \mathcal{L}(\mathcal{X} \Xi, Q=1)}, \quad E(\mathcal{X} \Xi) = - \sum_{q=1}^Q \sum_{n=1}^N h_q(\mathbf{x}_i) \ln h_q(\mathbf{x}_i)$	[10]

Split-Merge operations are criteria that are used to decide whether a component should be split or a merger of two components should occur. A split criterion could be based on the multivariate (MV) kurtosis, because a low or a high MV kurtosis value is an indication that a component should be split [21], [22]. However, the confidence intervals for multivariate kurtosis are accurate only asymptotically, i.e. when the number of sample measurement vectors tends to infinity [23]. A merge criterion of two components q, q' is the inner product [12]

$$J_{\text{merge}}(q, q') = [h_q(x_1), h_q(x_2), \dots, h_q(x_N)]^T [h_{q'}(x_1), h_{q'}(x_2), \dots, h_{q'}(x_N)]. \quad (11)$$

However, this criterion may not yield a merger of two non-Gaussian components to a single Gaussian one. It is recommended only for components with similar parameters, and in addition, the confidence intervals for this criterion have not been found yet. Sequences of split-merge operations can cause oscillations around a number of components, a fact which can increase the execution time.

The goal of this paper is to present a split technique that does not require any component merging. The proposed splitting criterion can be considered simply as a transformation of a D -dimensional space onto an one dimensional space. Subsequently, a univariate distribution test in the one dimensional space is derived. The transformation from many dimensions to one dimension is accomplished through the Mahalanobis distance of each sample measurement vector from the mean vector of a certain Gaussian

component, which will be called hereafter as Mahalanobis distance. The component where each sample measurement vector belongs to is found by an assignment that uses the unobserved variables. Such a criterion has been extensively used for assessing multivariate normality [24], [25], but it has not been explored yet as a plug-in criterion for splitting non-Gaussian components in EM. The Mahalanobis distance can be treated as a random variable (r.v.) that follows a certain beta pdf as it is proven in a lemma in [26]. Since the proof of this lemma is rare to find, it is rather complex, because it contains a series of theorems, and it can be easily confused with other proofs for several types of Mahalanobis distances, we revise a great part of the proof in the Appendix.

B. Alternative EM steps (2nd level)

Two methods that use different E- and M-Steps than the standard ones in order to escape from local optima have been reported, namely the deterministic annealing EM (DAEM) [13] and the component-wise EM (CEMM) [14]. In DAEM the E-Step is modified by a parameter $1/\beta \in [1, \infty)$, called temperature. Specifically, the unobservable variables $h_q(\mathbf{x}_i)$ are found by

$$h_q(\mathbf{x}_i) = \frac{[\pi_q f_q(\mathbf{x} | \Xi_q)]^\beta}{\sum_{q'=1}^Q [\pi_{q'} f_{q'}(\mathbf{x} | \Xi_{q'})]^\beta}. \quad (12)$$

As $1/\beta$ increases, $h_q(\mathbf{x}_i) \rightarrow 1/Q$, i.e. a sample measurement vector is more likely to belong to all components. Therefore, component parameters become similar and the chance to escape from a local optimum in the parameter space is high. A usual strategy is to set $\beta = 0.9$ until convergence of DAEM, to increase β by 0.05, i.e. $\beta \leftarrow \beta + 0.05$, and re-apply DAEM. The procedure stops when $\beta = 1$, where DAEM becomes actually the standard EM.

In the CEMM, the M-Step is altered as follows. In the r th iteration of EM, only the component indexed by $q = \text{mod}(r, Q) + 1$ is updated. This results to an M-Step that maximizes the log-likelihood function with a slower rate than that of the standard M-Step and it follows a longer path in the parameter space in order to converge to the final solution, which might yield better estimates of Ξ [14].

C. Initialization methods (1st level)

Several initialization methods for the parameters of each component can be found in the literature. In *random initialization* for GMMs, the component priors are equal to $1/Q$, the centers are randomly chosen sample measurement vectors, and the covariance matrices of the components denoted as \mathbf{S}_q are initialized as [9]

$$\mathbf{S}_q = \frac{1}{10D} \text{trace}(\mathbf{S}) \quad (13)$$

where \mathbf{S} is the covariance matrix of the entire \mathcal{X} . Initialization through k -means algorithm is also widely used [15], [13]. k -means, however, is itself sensitive to local optima of the parameter space and might yield a biased initialization. In *partial random* initialization, a component is randomly added in the GMM after convergence of the EM algorithm, and the parameters of the new component as well as the priors of the old components are refined with the EM algorithm. During this procedure the old component centers and old covariance matrices are kept fixed [16]. *Re-sampling techniques* use random [16], bootstrap [27] or cross-validation estimates [17] of the likelihood function by sampling the initial sample set in order to find the best initialization for EM, which, however, may not yield the global optimum of Ξ .

The contribution of this paper in the initialization level is in the initialization of two new components after splitting an old one. Splitting is accomplished either by a discriminant or by initializing the centers of the new components by setting them equal to the center of the old component. The MV kurtosis is used as a switch for deciding among the aforementioned split methods. A large multivariate kurtosis value of sample measurement vectors that belong to the old component indicates that this cluster of sample measurement vectors is an outcome of a leptokurtic distribution. Since we assume that only Gaussians exist in the mixture, the leptokurtic distribution can result when two or more Gaussian sources with common centers are present. Otherwise, if kurtosis value is small, then the cluster is an outcome of a platykurtic distribution. Platykurtic distributions could be obtained then, if two or more Gaussian sources with separate centers exist. Therefore, the old cluster is split by a discriminant. In the following, the outline of this paper is presented.

D. Outline

The outline of this paper is as follows. In Section II, the 5 steps of the proposed algorithm are described. The second and the third steps are detailed in separate sections. The second step of the proposed algorithm uses a *multivariate normality criterion* based on the Mahalanobis distance of each sample measurement vector from the component center to decide if a component should be split, as it is detailed in Section III. The third step of the proposed algorithm employs a central tendency criterion based on the expected MV kurtosis of the Gaussian density to initialize the centers of the two components during splitting, which is described in Section IV. Experimental results on artificially generated and real data-sets as well as comparisons against other EM variants are given in Section V. Finally, conclusions are drawn in Section VI.

II. ALGORITHM DESCRIPTION

The general idea of the proposed algorithm is to begin with a single cluster, split the cluster into two clusters, split the two clusters into three clusters and so on, until every cluster is the outcome of a single multivariate Gaussian source. The cluster to be split is found via a multivariate normality test based on the Mahalanobis distance of each sample measurement vector from the component center it belongs to. The cluster with the worst fit with respect to the Mahalanobis distance distribution is split into two clusters that will be called as new clusters hereafter. If the MV kurtosis of the old cluster is significantly large, the centers of the new clusters are set both equal to the old cluster center initially, otherwise the old cluster is split with a discriminant perpendicular to an axis.

Let the set of D -dimensional sample measurement vectors \mathfrak{X} be modeled by a mixture of Gaussian multivariate densities. That is, \mathfrak{X} is considered as the union of Q clusters $\mathcal{L}_q, q = 1, 2, \dots, Q$, and each cluster \mathcal{L}_q is a realization of Gaussian pdf \mathcal{G}_q . The goal is to find $\{\mathcal{G}_q\}_{q=1}^Q$. For readers' convenience, a flow chart of the algorithm is sketched in Figure 2. Let us assume the null hypothesis $H_0 = \{\mathcal{L}_q \sim \mathcal{G}_q\}_{q=1}^Q$, i.e. \mathfrak{X} is modeled by Q components \mathcal{G}_q where each component fits the cluster \mathcal{L}_q .

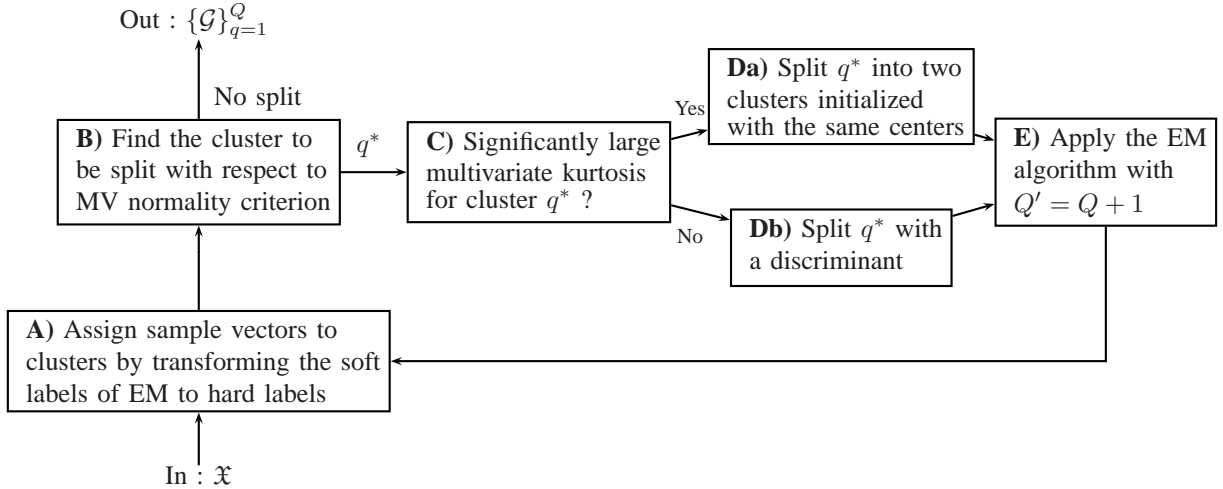


Fig. 2. Steps of the proposed algorithm.

Initially, we make the hypothesis that $H_0 = \mathcal{L}_1 \sim \mathcal{G}_1$, where $\mathcal{L}_1 = \mathfrak{X}$, i.e. \mathcal{L}_1 is the outcome of a single Gaussian source \mathcal{G}_1 . The parameters of \mathcal{G}_1 are the sample mean vector $\bar{\mathbf{x}}$ and sample dispersion matrix \mathbf{S} of \mathfrak{X} . Let $\mathcal{D}_{\mathcal{L}_1}$ be the criterion that measures the normality of cluster \mathcal{L}_1 . $\mathcal{D}_{\mathcal{L}_1}$ is the number of sample measurement vectors of \mathcal{L}_1 that are outside a proper confidence interval for the distribution of the Mahalanobis distance. $\mathcal{D}_{\mathcal{L}_1}$ will be analytically described in Section III. For the time being, it is

sufficient to test whether $\mathcal{D}_{\mathcal{L}_1} > (1 - \lambda)|\mathcal{L}_1|$, where $|\mathcal{L}_1|$ is the cardinality of sample measurement vectors that belong to \mathcal{L}_1 , in order to reject the hypothesis that \mathcal{L}_1 is the outcome of a single Gaussian source at $\lambda=99\%$ confidence level.

If $\mathcal{D}_{\mathcal{L}_1} > (1 - \lambda)|\mathcal{L}_1|$, the hypothesis $H_0 = \mathcal{L}_1 \sim \mathcal{G}_1$ is rejected. Accordingly, $\mathcal{L}_1 = \mathfrak{X}$ should be split into \mathcal{L}_1 and \mathcal{L}_2 clusters so that $\mathfrak{X} = \cup_{q=1}^2 \mathcal{L}_q$. We proceed to testing the hypothesis $H_0 = \{\mathcal{L}_q \sim \mathcal{G}_q\}_{q=1}^2$. The general hypothesis $H_0 = \{\mathcal{L}_q \sim \mathcal{G}_q\}_{q=1}^Q$ with $Q > 1$, is described next.

A) Assignment. Each sample measurement vector is assigned to a cluster $\mathcal{L}_1, \dots, \mathcal{L}_Q$ as follows. Let us assume that $h_q(\mathbf{x}_i)$ is the probability that a sample measurement vector belongs to component \mathcal{G}_q . $h_q(\mathbf{x}_i)$ are obtained by the EM algorithm after its convergence. Realizations ϱ_i , $i = 1, 2, \dots, N$ of a r.v. uniform in $[0, 1]$ are created. For every $i = 1, 2, \dots, N$,

$$\text{if } \varrho_i \in \left[\sum_{q'=1}^{q-1} h_{q'}(\mathbf{x}_i), \sum_{q'=1}^q h_{q'}(\mathbf{x}_i) \right], \text{ then } \mathbf{x}_i \in \mathcal{L}_q. \quad (14)$$

This assignment results to Gaussian distributed clusters, even if their components overlap. An example of 2 components is depicted in Figures 3(a) and 3(b).

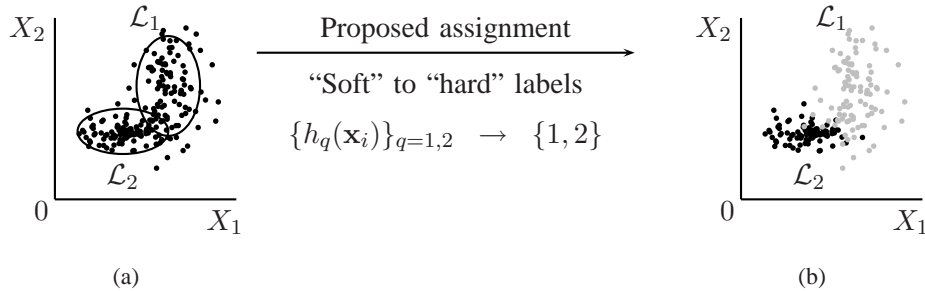


Fig. 3. a) Two Gaussian densities that overlap when the probability for a sample measurement vector \mathbf{x}_i belongs to component $q = 1, 2$ is $h_q(\mathbf{x}_i)$; b) Proposed hard assignment with the help of a random variable, so that clusters \mathcal{L}_1 and \mathcal{L}_2 are the outcome of two Gaussians.

B) Find the cluster to be split. Let q^* denote the index of the cluster to be split, where $q^* \in \{1, 2, \dots, Q\}$. Formally \mathcal{L}_{q^*} is the cluster that satisfies

$$q^* = \operatorname{argmax}_{q=1,2,\dots,Q} [\mathcal{D}_{\mathcal{L}_q} - (1 - \lambda)|\mathcal{L}_q|]. \quad (15)$$

If $\mathcal{D}_{\mathcal{L}_q} < (1 - \lambda)|\mathcal{L}_q|$, $\forall q = 1, 2, \dots, Q$, the algorithm stops because no cluster deviates from the MV normal distribution. Only one Gaussian is chosen to be split, because otherwise the algorithm starts splitting clusters that are modeled well by MV Gaussian densities. Such an example for $Q = 2$

components, namely \mathcal{G}_1 and \mathcal{G}_2 , is depicted in Figures 4(a), 4(b), and 4(c). An over-splitting case is shown in Figure 4(b), where both \mathcal{G}_1 and \mathcal{G}_2 are split. If \mathcal{G}_2 is only split, as shown in Figure 4(c), then the correct number of Gaussian components is found.

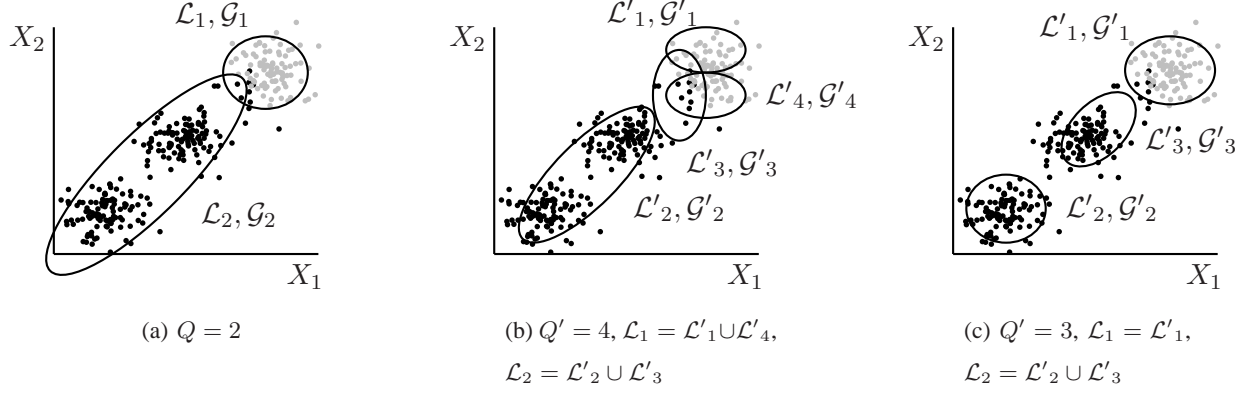


Fig. 4. a) GMM with $Q = 2$ components where $\mathcal{D}_{\mathcal{L}_2} - 0.01|\mathcal{L}_2| > \mathcal{D}_{\mathcal{L}_1} - 0.01|\mathcal{L}_1| > 0$; b) Both \mathcal{L}_1 and \mathcal{L}_2 are split; c) A better initialization for EM algorithm is obtained when \mathcal{L}_2 is split only.

C) Kurtosis switch: The splitting of \mathcal{L}_{q^*} into clusters \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$, with $Q' = Q + 1$ is performed either by a discriminant or by initializing both new component centers with the old component center. The choice of the splitting method depends on the value of MV kurtosis for cluster \mathcal{L}_{q^*} denoted as $K(\mathcal{L}_{q^*})$ [28]. A large $K(\mathcal{L}_{q^*})$ value indicates that \mathcal{L}_{q^*} is the outcome of a leptokurtic distribution. Since, only Gaussians exist in the mixture, the leptokurtic distribution could be the outcome of two or more Gaussian sources with common centers. An example of high MV kurtosis value is depicted in Figure 5, where \mathcal{L}_{q^*} is the outcome of two Gaussian sources with common centers. Figure 5(b) shows the initialization of EM, when \mathcal{L}_{q^*} is split by a discriminant, whereas in Figure 5(d) the initialization of EM with the centers of the new components initially set equal to the old center is presented. From the comparison of the GMMs in Figures 5(c) and 5(e), it can be inferred that the best GMM is found by initializing the centers of the new components with the center of the original component.

Let K_0 be the first-order moment of the kurtosis of the MV Gaussian distribution derived in Section IV. Splitting is done according to:

$$\text{if } K(\mathcal{L}_{q^*}) > K_0, \quad (16)$$

Da) initializing the centers of the new components with the old component center: \mathcal{L}_{q^*} is split into clusters \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$, where $Q' = Q + 1$, by initializing $\mu'_{q^*} \leftarrow \mu_{q^*}$ and $\mu'_{Q'} \leftarrow \mu_{q^*}$. Additionally, the priors π'_{q^*} and $\pi'_{Q'}$ of the new components are both set to one half of the initial a priori probability of

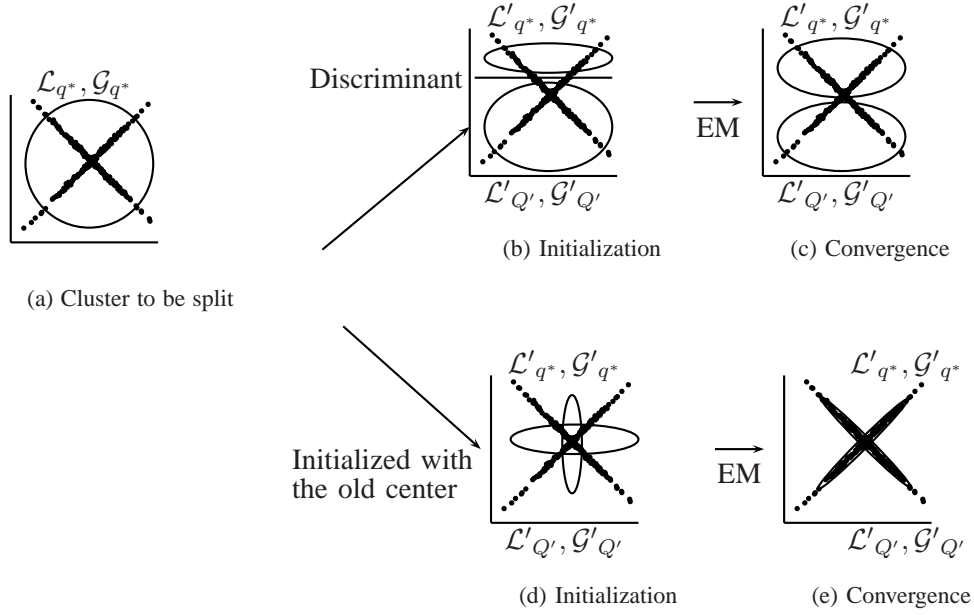


Fig. 5. Example of splitting cluster \mathcal{L}_{q^*} .

the cluster, i.e. $\frac{|\mathcal{L}_{q^*}|}{2|\mathcal{X}|}$. The covariance matrices $\mathbf{S}'_{q^*}, \mathbf{S}'_{Q'}$ are randomly initialized. A random initialization of covariance matrices is done by setting $\mathbf{S}'_{q^*}, \mathbf{S}'_{Q'}$ equal to two different $D \times D$ diagonal matrices, respectively. The diagonal elements of each matrix are realizations of the r.v. s , where

$$s^2 \frac{2D(|\mathcal{L}_{q^*}| - 1)}{||\mathbf{S}_{q^*}||} \sim \chi^2_{|\mathcal{L}_{q^*}| - 1}. \quad (17)$$

The random initialization of the covariance matrix stems from the theorem stating that marginal variance should follow the χ^2 distribution with $|\mathcal{L}_{q^*}| - 1$ degrees of freedom. We used (17) which produces different covariance matrices instead of (13) that results to identical covariance matrices in order to avoid creating two new components, which would have the same covariance matrices on the top of the same centers and the same a priori probabilities. If the two new components, created after a split, had the same covariance matrices, then according to (5)-(8), the component parameters Ξ^{r+1} would be equal to Ξ^r , i.e. Ξ would not be optimized. Otherwise, a

Db) discriminant is applied: That is, \mathcal{L}_{q^*} is split into clusters \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$, where $Q' = Q + 1$, by a discriminant hyperplane found with respect to marginal statistics. The discriminant hyperplane is the value of vector $\mathbf{x}_{i^*} \in \mathcal{L}_{q^*}$ along axis X_{d^*} found by

$$x_{i^*d^*} = \underset{\substack{d = 1, 2, \dots, D \\ i = 1, 2, \dots, |\mathcal{L}_{q^*}|}}{\operatorname{argmax}} F_{X_d}(x_{id}) - \hat{F}_{X_d}(x_{id}), \quad (18)$$

where $F_{X_d}(x_{id})$ is the theoretical marginal Gaussian cumulative distribution function (cdf) with parameters estimated by the marginal sample mean and variance, and $\hat{F}_{X_d}(x_{id})$ is the empirical marginal cdf on X_d -axis calculated from the *mass function* [29]. The theoretical marginal Gaussian cdf is found via the error function. The hyperplane $x_{i^*d^*}$ is perpendicular to X_{d^*} -axis and has the property of dividing a cluster into two separate clusters. For example, in Figure 6(a), $x_{i^*d^*}$ is chosen as value $\mathbf{x}_{i^*} \in \mathfrak{X}$ onto axis X_2 , because, as it is seen from the comparison of Figures 6(b) and 6(c) the highest distance $F_{X_d}(x_{id}) - \hat{F}_{X_d}(x_{id})$ is observed for $d = 2$. After splitting \mathcal{L}_{q^*} into clusters \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$, their centers, sample dispersion matrices, and priors are used to initialize the EM algorithm. By making the initialization as in Figure 6(d), the EM converges to the most descriptive GMM shown in Figure 6(e).

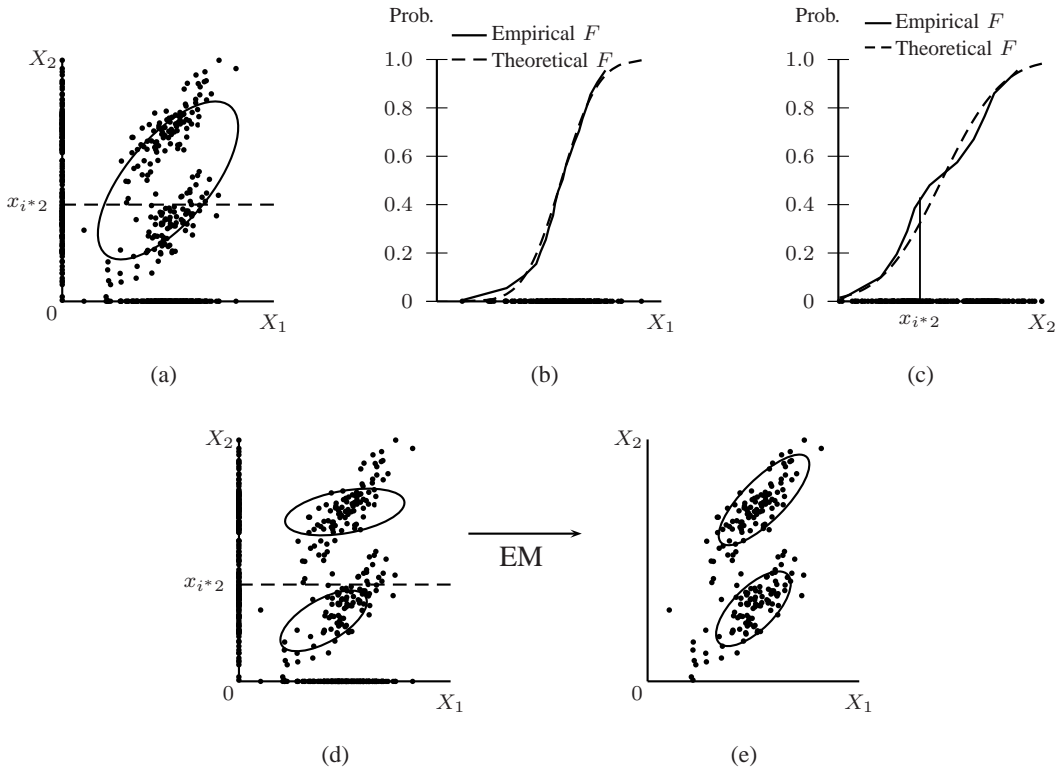


Fig. 6. a) Modeling a set of 2D sample measurement vectors, b) comparison between the marginal normal cdf and empirical marginal cdf for X_1 , c) the same for X_2 , d) splitting by a discriminant, and e) GMM after EM convergence.

E) Apply the EM algorithm. The EM algorithm refines the GMM model iteratively, with initial $\Xi'^1 = \{\pi_q'^1, \mu_q'^1, \mathbf{S}_q'^1\}_{q=1}^{Q'}$ set as

$$\pi_q'^1 = \frac{|\mathcal{L}'_q|}{N}, \quad q = 1, 2, \dots, Q', \quad \text{where } N = |\mathfrak{X}|, \quad (19)$$

$$\mu_q'^1 = \frac{1}{|\mathcal{L}'_q|} \sum_{\mathbf{x}_i \in \mathcal{L}'_q} \mathbf{x}_i, \quad \text{and} \quad (20)$$

$$\mathbf{S}_q'^1 = \frac{1}{|\mathcal{L}'_q| - 1} \sum_{\mathbf{x}_i \in \mathcal{L}'_q} (\mathbf{x}_i - \boldsymbol{\mu}_q'^1)(\mathbf{x}_i - \boldsymbol{\mu}_q'^1)^T. \quad (21)$$

The EM algorithm stops when

$$|\mathcal{L}(\mathbf{x} | \boldsymbol{\Xi}'^{r+1}) - \mathcal{L}(\mathbf{x} | \boldsymbol{\Xi}'^r)| < 10^{-5} |\mathcal{L}(\mathbf{x} | \boldsymbol{\Xi}'^r)|. \quad (22)$$

Obviously (22) is a heuristic method to find the local maximum of the log-likelihood function used many times in literature [9], [16]. It has been employed in the experiments reported in Section V-B. The absolute values in (22) are necessary, because $\mathcal{L}(\mathbf{x} | \boldsymbol{\Xi})$ can be negative, since it involves the logarithmic operator.

Steps (A) to (E) are repeated with the newly found parameters, i.e. $\mathcal{L}_q \leftarrow \mathcal{L}'_q$, $\mathcal{G}_q \leftarrow \mathcal{G}'_q$ for $q = 1, 2, \dots, Q'$, and $Q \leftarrow Q'$. The algorithm stops when no cluster diverges from the MV Gaussian. The proposed algorithm is summarized in Figure 7.

Input is $\mathcal{L}_1 \leftarrow \mathcal{X}$ and initial hypothesis is $H_0 = \mathcal{L}_1 \sim \mathcal{G}_1$. Set number of components $Q \leftarrow 1$.

- Estimate the MV normality criterion $\mathcal{D}_{\mathcal{L}_1}$ according to algorithm summarized in Figure 9.
- If $\mathcal{D}_{\mathcal{L}_1} < (1 - \lambda)|\mathcal{L}_1|$ stop, else split \mathcal{L}_1 according to steps (C-D).

A) Assign sample sample measurement vectors to clusters

- Create realizations ϱ_i , $i = 1, 2, \dots, N$ of a r.v. uniform in $[0, 1]$ and apply (14).

B) Test whether $H_0 = \{\mathcal{L}_q \sim \mathcal{G}_q\}_{q=1}^Q$.

- Find $q^* = \underset{q=1,2,\dots,Q}{\operatorname{argmax}} [\mathcal{D}_{\mathcal{L}_q} - (1 - \lambda)|\mathcal{L}_q|]$.
- If $\mathcal{D}_{\mathcal{L}_{q^*}} < (1 - \lambda)|\mathcal{L}_{q^*}|$ **stop**, else split \mathcal{L}_{q^*} according to steps (C-D).

C-D) Split \mathcal{L}_{q^*} into \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$, where $Q' \leftarrow Q + 1$.

C) Find $K(\mathcal{L}_{q^*})$ from (33), and K_0 from (35).

Da) If $K(\mathcal{L}_{q^*}) > K_0$, then initialize $\boldsymbol{\mu}'_{q^*} \leftarrow \boldsymbol{\mu}_{q^*}$ and $\boldsymbol{\mu}'_{Q'} \leftarrow \boldsymbol{\mu}_{q^*}$. Covariance matrices of new components $\mathbf{S}'_{q^*}, \mathbf{S}'_{Q'}$ are randomly initialized according to (17). Also set $\pi'_{q^*} \leftarrow \frac{|\mathcal{L}_{q^*}|}{2|\mathcal{X}|}$ and $\pi'_{Q'} \leftarrow \frac{|\mathcal{L}_{q^*}|}{2|\mathcal{X}|}$.

Db) Else if $K(\mathcal{L}_{q^*}) < K_0$, split \mathcal{L}_{q^*} into \mathcal{L}'_{q^*} and $\mathcal{L}'_{Q'}$ by the discriminant found with (18).

- The remaining clusters remain intact, i.e. $\mathcal{L}'_q \leftarrow \mathcal{L}_q$ for $q = 1, 2, \dots, q^* - 1, q^* + 1, \dots, Q$.

E) Initialize EM with $\{\mathcal{L}'_q\}_{q=1}^{Q'}$, and repeat E- and M-Steps until convergence according to (22).

- Refine GMM by setting $\mathcal{L}_q \leftarrow \mathcal{L}'_q$, $\mathcal{G}_q \leftarrow \mathcal{G}'_q$, for $q = 1, 2, \dots, Q'$, and $Q \leftarrow Q'$ and go to step (A).

Fig. 7. Proposed clustering algorithm based on EM.

III. HYPOTHESIS TESTING FOR MV NORMALITY WITH RESPECT TO MAHALANOBIS DISTANCE

A process to establish a hypothesis that a random vector (R.V.) $\underline{x} = [X_1, X_2, \dots, X_D]^T$ is distributed according to the multivariate Gaussian distribution is presented. Let $\mathfrak{X} = \{\mathbf{x}_i\}_{i=1}^N$ be a set of N sample measurement vectors $\mathbf{x}_i \in \mathbb{R}^D$ of the R.V. \underline{x} . For example, N sample measurement vectors of a D -dimensional R.V. are depicted in Figure 8, where D is limited to 2. The Mahalanobis distance of $\mathbf{x}_i \in \mathfrak{X}$ from the center of \mathfrak{X} is defined as

$$r_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}). \quad (23)$$

The empirical cdf of the r.v. R_i admitting values r_i , denoted as $\hat{F}_{R_i}(r_i)$, is found via the mass function, i.e. by sorting $\{r_i\}_{i=1}^N$ in ascending order and by letting $\hat{F}_{R_i}(r_i) = i/N$. Let $F_{R_i}(r_i)$ be the theoretical cdf of R_i given the mean vector $\bar{\mathbf{x}}$ and the sample dispersion matrix \mathbf{S} of \mathfrak{X} , which is revised in the Appendix. If N_{r_i} denotes the number of sample measurement vectors inside the r_i -equiprobable ellipse, then it can be inferred that N_{r_i} is a binomial r.v. with parameters N and $F_{R_i}(r_i)$, i.e.

$$P(N_{r_i} = k) = \binom{N}{k} (F_{R_i}(r_i))^k (1 - F_{R_i}(r_i))^{N-k}, \quad (24)$$

because $F_{R_i}(r_i)$ is also the probability of having a sample measurement vector inside the ellipse with Mahalanobis distance equal to r_i .

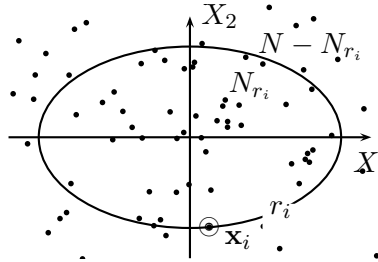


Fig. 8. r_i defines an ellipse that separates the sample set into two populations.

Let $k_{i;\lambda}^l \in [0, N]$ be the low confidence limit of N_{r_i} at $100\lambda\%$ confidence level. Let $k_{i;\lambda}^h \in [0, N]$ be the high confidence limit of N_{r_i} for $i = 1, 2, \dots, N$. The confidence limits should satisfy

$$\sum_{k=k_{i;\lambda}^h}^N \binom{N}{k} (F_{R_i}(r_i))^k (1 - F_{R_i}(r_i))^{N-k} = \sum_{k=0}^{k_{i;\lambda}^l} \binom{N}{k} (F_{R_i}(r_i))^k (1 - F_{R_i}(r_i))^{N-k} = \frac{1 - \lambda}{2}, \quad (25)$$

where $\lambda \in \{0.90, 0.95, 0.99\}$ in most cases [29]. Starting with the results in [29], we revise the algorithm to find the confidence interval $(k_{i;\lambda}^l, k_{i;\lambda}^h)$ for a binomial r.v., subsequently. The novelty in this section is the derivation of $(k_{i;\lambda}^l, k_{i;\lambda}^h)$, which is the confidence interval for the number of sample measurement vectors inside the ellipse defined by r_i at $100\lambda\%$ level of confidence.

First, if N is large enough and $F_{R_i}(r_i)$ is neither near 0 nor near 1, i.e.,

$$NF_{R_i}(r_i)(1 - F_{R_i}(r_i)) \gg 1, \quad (26)$$

according to the DeMoivre-Laplace theorem the binomial distribution can be approximated by a Gaussian distribution with mean $NF_{R_i}(r_i)$ and variance $NF_{R_i}(r_i)(1 - F_{R_i}(r_i))$ [29]. A typical value for this assumption is $NF_{R_i}(r_i)(1 - F_{R_i}(r_i)) > 25$ [30]. So,

$$(k_{i;\lambda}^l, k_{i;\lambda}^h) = \left([NF_{R_i}(r_i) - z_\lambda \sqrt{2NF_{R_i}(r_i)(1 - F_{R_i}(r_i))}], [NF_{R_i}(r_i) + z_\lambda \sqrt{2NF_{R_i}(r_i)(1 - F_{R_i}(r_i))}] \right), \quad (27)$$

where $[]$ denotes the closest integer to the number inside delimiters, and z_λ equals to 1.16, 1.39, 1.82, for $\lambda = 0.9, 0.95, 0.99$, respectively.

Second, if (26) is violated, then the confidence interval $(k_{i;\lambda}^l, k_{i;\lambda}^h)$ is estimated by

$$k_{i;\lambda}^l = \underset{k_1=0,1,\dots,N}{\operatorname{argmin}} \left| \sum_{k=0}^{k_1} \binom{N}{k} (F_{R_i}(r_i))^k (1 - F_{R_i}(r_i))^{N-k} - \frac{1-\lambda}{2} \right|, \quad (28)$$

$$k_{i;\lambda}^h = \underset{k_2=N,\dots,1,0}{\operatorname{argmin}} \left| \sum_{k=k_2}^N \binom{N}{k} (F_{R_i}(r_i))^k (1 - F_{R_i}(r_i))^{N-k} - \frac{1-\lambda}{2} \right|. \quad (29)$$

The hypothesis test should validate if

$$N_{r_i} \in (k_{i;\lambda}^l, k_{i;\lambda}^h) \Rightarrow \frac{N_{r_i}}{N} \in \left(\frac{k_{i;\lambda}^l}{N}, \frac{k_{i;\lambda}^h}{N} \right) \Rightarrow \hat{F}_{R_i}(r_i) \in \left(\frac{k_{i;\lambda}^l}{N}, \frac{k_{i;\lambda}^h}{N} \right), \quad (30)$$

$\forall i = 1, 2, \dots, N$. So we formulate the following null hypothesis

$H_0 = \mathfrak{X} \sim \mathcal{G}$: The hypothesis \mathfrak{X} stems from \mathcal{G} is accepted at $100\lambda\%$ confidence level if

$$\hat{F}_{R_i}(r_i) \in \left(\frac{k_{i;\lambda}^l}{N}, \frac{k_{i;\lambda}^h}{N} \right) \text{ for at least } \lambda|\mathfrak{X}| \text{ out of } |\mathfrak{X}| \text{ times.}$$

For example, a set of sample measurement vectors $\mathfrak{X} = \{(x_{i1}, x_{i2})^T\}_{i=1}^{200}$ that stems from a mixture of two Gaussians is artificially generated and plotted in Figure 10. Let one Gaussian be fitted onto \mathfrak{X} . The ellipse corresponds to $r = 1.2$. Let us assume that r_i are sorted in ascending order. The MV test is applied to test null hypothesis $H_0 = \mathfrak{X} \sim \mathcal{G}$. In Figure 11, the empirical cdf of r_i , i.e. $\hat{F}_{R_i}(r_i) = i/N$, is plotted and compared against its confidence intervals estimated from (27)-(29). $\hat{F}_{R_i}(r_i)$ is significantly lower than the theoretical one $F_{R_i}(r_i)$, when $r < 1.2$. That is, less sample measurement vectors than expected are inside the corresponding ellipse. The MV normality criterion value $\mathcal{D}_{\mathfrak{X}}$ counts how many times $\hat{F}_{R_i}(r_i)$ falls outside $(\frac{k_{i;\lambda}^l}{N}, \frac{k_{i;\lambda}^h}{N})$, i.e.

$$\mathcal{D}_{\mathfrak{X}} = \sum_{\hat{F}_{R_i}(r_i) - \frac{k_{i;\lambda}^h}{N} > 0 \text{ or } \frac{k_{i;\lambda}^l}{N} - \hat{F}_{R_i}(r_i) < 0} 1. \quad (31)$$

If $\mathcal{D}_{\mathfrak{X}} > (1 - \lambda)N$, then $H_0 = \mathfrak{X} \sim \mathcal{G}$ is rejected at $100\lambda\%$ significance level. For example, if $\lambda = 0.95$ and $N = 100$, $\mathcal{D}_{\mathfrak{X}}$ should be greater than 5 in order to reject $H_0 = \mathfrak{X} \sim \mathcal{G}$.

The value of λ is chosen according to the value of N due to quantization, i.e. since $\mathcal{D}_{\mathfrak{X}} \in \{0, 1, 2, \dots, N\} \Rightarrow \mathcal{D}_{\mathfrak{X}}/N \in \{0, 1/N, 2/N, \dots, 1\}$, so $\lambda \in \{0, 1/N, 2/N, \dots, 1\}$. If N is small (e.g. if $N = 20$), then $\lambda \in \{0, 0.05, 0.1, \dots, 1\}$, so λ can not be 0.99. In order to avoid such discrepancies, we propose

$$\lambda = \begin{cases} 0.99 & \text{if } N \geq 100, \\ 0.95 & \text{if } 20 \leq N < 100, \\ 0.9 & \text{if } 10 \leq N < 20. \end{cases} \quad (32)$$

If $N < 10$, \mathfrak{X} is not split, because according to (32) the significance level λ should be below 0.9. The proposed algorithm for testing whether a set of measurement vectors stems from a single multivariate Gaussian component is summarized in Figure 9.

- 1) Estimate $r_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$ for each $i = 1, 2, \dots, N$;
- 2) Sort $\{r_i\}_{i=1}^N$ in ascending order, and set $\hat{F}_{R_i}(r_i) = i/N$;
- 3) Evaluate the confidence intervals $(k_{i;\lambda}^l, k_{i;\lambda}^h)$ using (27), (28), and (29).
- 4) The hypothesis $H_0 = \mathfrak{X} \sim \mathcal{G}$ that the sample set \mathfrak{X} stems from the multivariate Gaussian \mathcal{G} is rejected at $100\lambda\%$ confidence level if $\mathcal{D}_{\mathfrak{X}} > (1 - \lambda)N$, where $\mathcal{D}_{\mathfrak{X}}$ and λ are given by (31) and (32), respectively.

Fig. 9. The MV normality criterion.

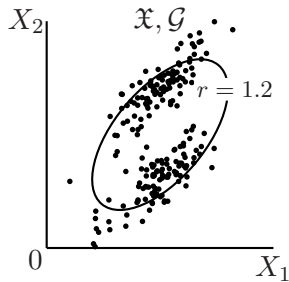


Fig. 10. A set of 2D sample measurement vectors.

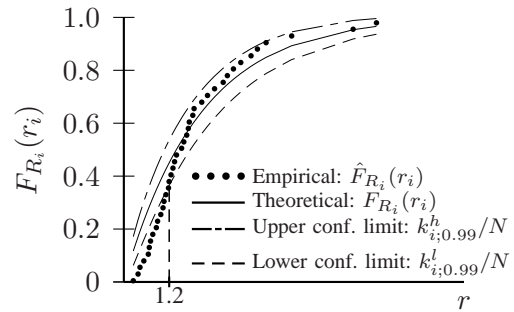


Fig. 11. Multivariate normality criterion for the set of sample measurement vectors shown in Figure 10.

The MV kurtosis and the expected MV kurtosis for the Gaussian case, that are used in the proposed EM algorithm, will be explained next.

IV. MULTIVARIATE KURTOSIS TEST

The multivariate (MV) kurtosis of a set of realizations $\mathfrak{X} = \{\mathbf{x}_i\}_{i=1}^N$ of the D -dimensional R.V. $\underline{x} = [X_1, X_2, \dots, X_D]^T$ is defined by K. Mardia as [28]

$$K(\mathfrak{X}) = \frac{1}{N} \sum_{i=1}^N r_i^2, \quad (33)$$

where r_i is the Mahalanobis distance estimated by (23).

Multivariate kurtosis is a measure of the peakedness of a cluster [31]. It is experimentally found that it can be used to detect if a cluster \mathfrak{X} is the result of two or more MV Gaussian sources with common centers. This observation is supported by the following reasoning: Large kurtosis indicates that \mathfrak{X} stems from a leptokurtic distribution, whereas a low kurtosis denotes that \mathfrak{X} stems from a platykurtic distribution. Since it is assumed that Gaussian sources only underlie the sample measurement vectors, a leptokurtic distribution happens only if two or more Gaussian densities share a common center, whereas a platykurtic distribution happens when the distance between the centers of the underlying Gaussian sources is large. From (33), it is evident that the domain of $K(\mathfrak{X})$ is $(0, \infty)$. Let us assume that a MV Gaussian density has expected kurtosis (or first-order moment) K_0 and $[K_{0;0.025}, K_{0;0.975}]$ is its confidence interval at 95% level of significance. By definition the order of these values is

$$0 < K_{0;0.025} < K_0 < K_{0;0.975} < \infty. \quad (34)$$

Three cases exist, namely

- H_0 : if $K(\mathfrak{X}) \in [K_{0;0.025}, K_{0;0.975}]$, \mathfrak{X} is distributed according to the MV Gaussian pdf;
- H_1 : if $K(\mathfrak{X}) \in (0, K_{0;0.025})$, then \mathfrak{X} is platykurtic;
- H'_1 : if $K(\mathfrak{X}) \in (K_{0;0.975}, \infty)$, \mathfrak{X} is leptokurtic.

We wish to establish if H'_1 is true or not. The following mathematical reasoning is applied. By using the multivariate normality test based on Mahalanobis distance described in Section III, the necessary information to establish whether H_0 is valid or not is obtained. If H_0 is not valid, either H_1 or H'_1 will be valid. To check which of the alternatives H_1 or H'_1 is valid, we examine whether $K(\mathfrak{X}) > K_0$ holds or not. If $K(\mathfrak{X}) > K_0 > K_{0;0.025}$ is valid, then also $K(\mathfrak{X}) > K_{0;0.025}$ is valid. So H_1 can not be valid, and by reduction ad absurdum H'_1 should be valid. Thus, it is established that \mathfrak{X} is leptokurtic without having to estimate $K_{0;0.975}$. K_0 can be easily estimated. To the opposite, only approximations for $K_{0;0.975}$ exist, when N is great [23]. More specifically, Mardia estimated K_0 [23]. According to the following derivations we found that K_0 , as is estimated by Mardia, is inaccurate for small N . Therefore we propose a better estimate than that of Mardia.

Theorem 1: The first-order moment of MV kurtosis is

$$K_0 = E(K) = \left(1 - \frac{1}{N}\right)^2 \frac{N-1}{N+1} D(D+2). \quad (35)$$

Proof: Let us assume that r_i are realizations of r.v.s. R_i . By applying the average operator to both sides of (33), we obtain

$$E(K) = \frac{1}{N} E\left(\sum_{i=1}^N R_i^2\right). \quad (36)$$

It is known from Appendix that R_i are identically distributed r.v.s. according to

$$\frac{N}{(N-1)^2} R_i \sim f_{Beta}(r_i \mid \frac{D}{2}, \frac{N-D-1}{2}), \quad i = 1, 2, \dots, N, \quad (37)$$

and it is also known that if r.v. $X \sim f_{Beta}(x \mid a, b)$, then [30]

$$E(X^M) = \prod_{m=0}^{M-1} \frac{a+m}{a+b+m}. \quad (38)$$

So from (37) and (38), it can be inferred that,

$$E\left(\frac{N^M}{(N-1)^{2M}} R_i^M\right) = \prod_{m=0}^{M-1} \frac{\frac{D}{2} + m}{\frac{N-1}{2} + m}, \quad (39)$$

or

$$E(R_i^M) = \frac{(N-1)^{2M}}{N^M} \prod_{m=0}^{M-1} \frac{D+2m}{N-1+2m}, \quad \forall i = 1, 2, \dots, N, \quad (40)$$

for all orders $M = 1, 2, \dots$

From (40), it is deduced that

$$E(R_i^M) = E(R_j^M) \quad \text{if } j \neq i, \quad (41)$$

for $i, j = 1, 2, \dots, N$ and $M = 1, 2, \dots$

By using (41), (36) becomes

$$E(K) = \frac{1}{N} \sum_{i=1}^N E(R_i^2) = E(R_i^2). \quad (42)$$

For $M = 2$, (40) yields (35). ■

The usefulness of the proposed estimator (35) is demonstrated in the following lines.

V. EXPERIMENTAL RESULTS

Experiments are divided into three sets. Experimental evidence to validate the accuracy of (35) is included in subsection V-A. Comparisons of the proposed GMM method against other GMM variants are performed in subsection V-B, and finally the initialization offered by the proposed MV kurtosis test in typical clustering cases is demonstrated in subsection V-C.

A. Experiments on the proposed estimate of expected MV kurtosis (35)

By comparing the proposed estimate (35) of MV kurtosis to

$$E(K) = \frac{N-1}{N+1}D(D+2), \quad (43)$$

derived by Mardia [23, eq. 3.16], it is easily seen that the two estimates differ by the factor $(1 - \frac{1}{N})^2$. As $\lim_{N \rightarrow \infty} (1 - \frac{1}{N})^2 = 1$, the difference becomes negligible. In Figure 12, the proposed estimate (35) of the average of multivariate kurtosis for feature dimension $D = 2, 5$, and 10 , and varying N , is compared against the standard estimate (43), and the empirical estimate found with Monte-Carlo repetitions. The latter is found by averaging the kurtosis $K(\mathcal{X})$ for 1000 artificially generated sets \mathcal{X} . As can be seen in Figures 12(a), (b), and (c), the proposed estimate (35) is closer to the empirical one for all values of N , whereas the one suggested by Mardia (43) is accurate only for large N .

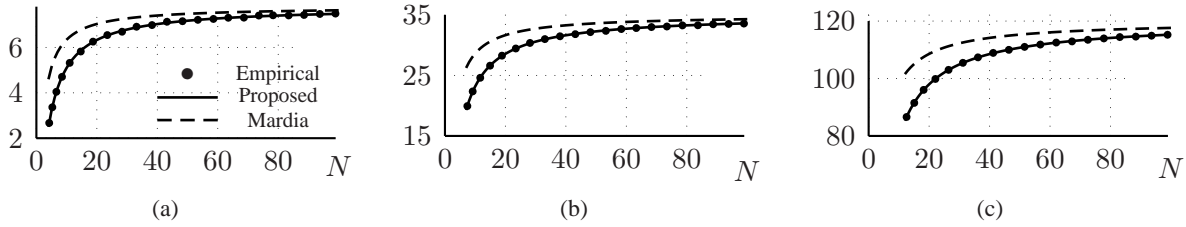


Fig. 12. The first-order moment of the multivariate kurtosis $E(K)$ of a MV Gaussian distributed cluster with respect to the number of sample measurement vectors N , and for feature dimension (a) $D = 2$ (b) $D = 5$, and (c) $D = 10$.

B. Comparison of the proposed method for GMM against other GMM methods.

The proposed algorithm is compared against 7 other EM variants according to 4 evaluation criteria for 5 data-sets over 1000 repetitions of the same experiment.

Data-sets: Three artificially generated data-sets and two real data-sets were used. The artificially generated data-sets are proposed as benchmark data for testing EM variants in other investigations [13], [9], [11]. The parameters for each of the three artificial generated data-sets as well as one realization of each data-set can be found in Figures 13(a), 13(b), and 13(c), respectively. Set \mathcal{A} is composed of few well separated components. Set \mathcal{B} is a mixture of few heavily overlapped components with different priors. Set \mathcal{C} is a set of many partially overlapping components with equal priors.

The two real data sets are utterances extracted from the Speech Under Simulated and Actual Stress data collection [32]. The utterances are 35 isolated words such as “break”, “go”, “one”, expressed from 9 male military persons in a studio environment. Each utterance is expressed two times by each speaker.

The first real data-set, denoted as Set \mathcal{D} , contains 1890 utterances equally separated to 3 speech styles (classes), namely slow, neutral, and fast. Each style is modeled as a mixture of Gaussians, where feature vectors extracted contain 5 features, namely the maximum duration of pitch contour plateaux at maxima, the median of durations for the rising slopes of pitch contour, the median of durations for the falling slopes of pitch contour, the maximum energy value, and the energy in the band 1-2.8 kHz normalized by the duration of each utterance. Set \mathcal{E} is the second real data-set that contains a total of 1890 utterances in anger, neutral, and soft speech styles. A two dimensional feature vector was used consisting of the energy values within the falling slopes of energy contours and the energy in the frequency band 3.5-3.95 kHz.

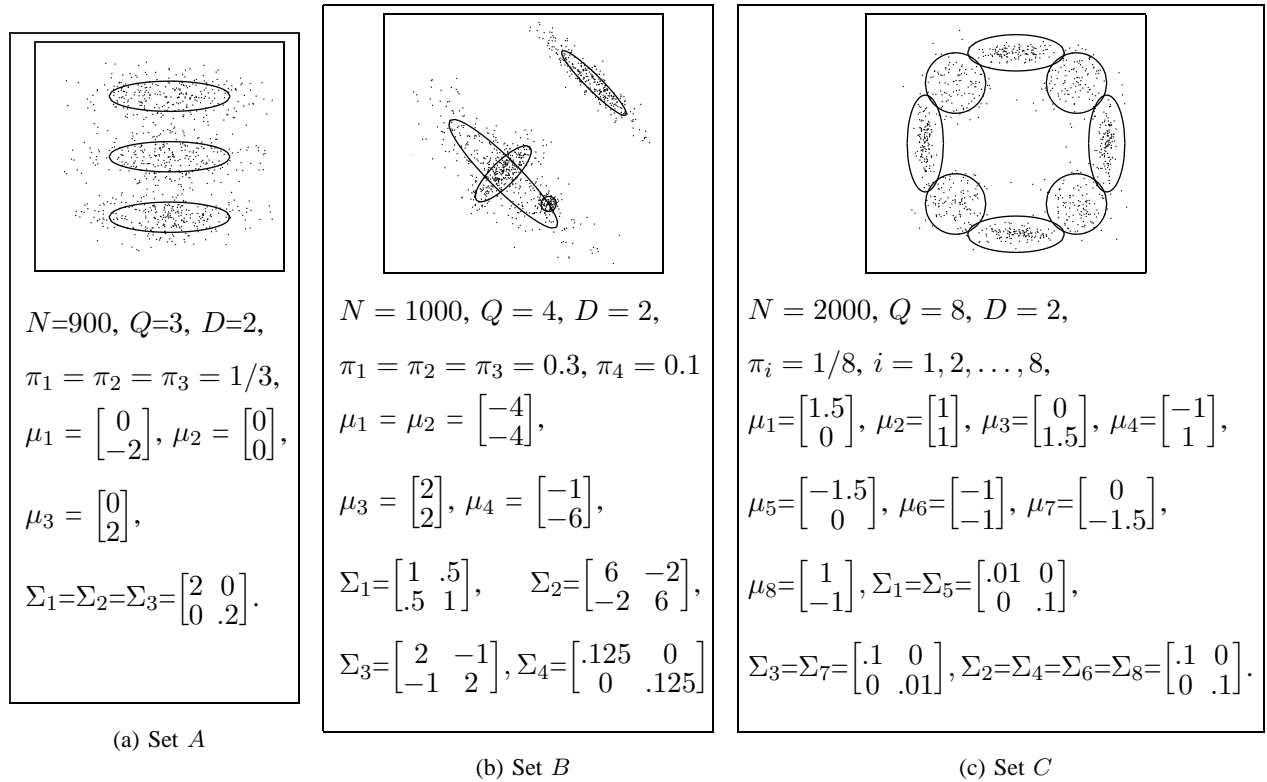


Fig. 13. Three artificially generated data-sets.

Methods compared: According to the categorization in Figure 1, we term the EM variants according to the following template: “3rd level technique - 2nd level technique - 1st level technique”. For example, “Forward MDL-EM-Random” is the EM variant that employs the forward logic with the MDL criterion to estimate the number of components, and the standard EM steps to refine a randomly initialized GMM. By

using the aforementioned terminology seven EM variants, those listed in the second column of Table II, are included in our comparative study. The convergence of each EM variant is judged according to (22). In 5th, 6th, and 7th methods, we chose to estimate the number of components Q with the MDL criterion, because the authors do not define a method to estimate it. In methods 1 to 6, the random initialization is preferred than k -means, so that results are comparable.

Evaluation criteria: The comparison is performed according to the following criteria:

- *Correctness* (in %): Correctness is the ratio of the times a correct GMM is found in 1000 Monte-Carlo repetitions. In each Monte-Carlo repetition of the experiment, a new realization of the data-set is generated. *Correctness is evaluated only for artificially generated data-sets*, because the true underlying Gaussian sources in real data-sets are unknown.
- *Prediction error* (in %): The classification error of the Bayes classifier when each class conditional pdf of real data is modeled by a mixture of Gaussians in 1000 cross-validation repetitions, where 90% of the available data was used for designing the GMMs and 10% for evaluating the prediction error [33]. *Prediction error is used instead of correctness in order to evaluate the performance of EM methods for real data*. The confidence intervals for the prediction error are estimated from the variance of prediction error in 1000 cross-validation repetitions, where it is assumed that the prediction error follows the Gaussian distribution.
- *Average number of EM iterations*: It is the average number of EM iterations required for an EM method to converge in 1000 Monte-Carlo repetitions. *It is not used in real data-sets, where the true model is unknown*.
- *Average execution time* (in sec): It is the average execution time measured over 1000 Monte-Carlo repetitions for artificially generated data or from 1000 cross-validation repetitions for real data. *It is more indicative about the computational needs of each EM method than the average number of EM steps*. The experiments are conducted on a PC with Pentium 4 CPU at 3 GHz and 1 Gb RAM at 400 MHz, by using Matlab 7.1.

From the results for artificially generated data presented in Table II, it can be inferred that the proposed method is the most accurate one for Set \mathcal{A} with 91.8% correctness, while maintaining the second lowest execution time, i.e. 0.72 sec. The 3rd and 6th methods follow with 85% and 77.9%, respectively. However, method 6 with 846 iterations is rather slow, which is due to the temperature parameter involved, which takes three values, namely $\frac{1}{0.9}$, $\frac{1}{0.95}$, 1. High execution time is observed for method 4, because it begins with $Q = 28$ components to reach finally $Q = 3$. The lowest execution time has been measured for

TABLE II
COMPARISON WITH OTHER EM VARIANTS FOR ARTIFICIAL DATA

		Correctness (%)			Average EM iterations			Average execution time (sec)		
#	Method/Data-set	\mathcal{A}	\mathcal{B}	\mathcal{C}	\mathcal{A}	\mathcal{B}	\mathcal{C}	\mathcal{A}	\mathcal{B}	\mathcal{C}
1	Forward MDL-EM-Random	71.3	2.2	23.1	329	96	512	1.13	0.34	7.45
2	Forward AIC-EM-Random	76.6	4.4	22.5	380	107	532	1.29	0.37	7.07
3	Forward ICL-EM-Random [8]	85	32.1	22.1	381	208	546	1.25	0.87	7.26
4	Backward MDL ₂ -EM-Random [9]	71.2	57.7	86.8	606	452	685	5.86	5.25	17.94
5	Forward MDL-CEMM-Random [14]	71.1	0	15.9	324	41	1663	0.83	0.12	14.96
6	Forward MDL-DAEM-Random [13]	77.9	0.5	0	846	196	245	3.16	0.72	1.74
7	Forward MDL-EM-Partial Random [16]	59.2	57.1	96.3	52	42	164	0.29	0.31	2.09
8	Split-EM-Discriminant (Proposed)	91.8	65.9	77.8	80	115	267	0.72	1.27	5.67

method 7. However, the partial random initialization leads to local optima of the EM algorithm and correctness drops to 59.2%.

For Set \mathcal{B} , methods 1, 2, 5, and 6 find only two components instead of four. This is due to the fact that the parsimonious criteria yield local minima with respect to Q , that are confused with the global minimum. A solution would be to inspect all possible Q . This strategy is followed by method 4, which however is rather slow, as it is seen from its execution time. It is confirmed that ICL [8] and MDL₂ [9] criteria employed in methods 3 and 4, respectively, are not so sensitive to local minima as MDL and AIC criteria used in methods 1, 2, 5 and 6. Correctness for each EM method drops for this set, since the prior of the fourth component is small, i.e. 0.1, and greatly overlaps with another component. The proposed method achieved 65.9% correctness, the highest one for this set, but its execution time is 1.27 sec, which is rather long compared to 0.31 sec of method 7.

Methods 4 and 7 achieved 86.8% and 96.3% correctness against 77.8% achieved by the proposed method for Set \mathcal{C} . Method 4, however requires 17.94 sec execution time, which is three times bigger than 5.67 sec needed for by the proposed method. For this data-set, method 7 has shown the highest accuracy with 96.3% and the lowest execution time at 2.09 sec.

The prediction error and execution time results for real data-sets are presented in Table III. In addition, the prediction error when the design set is used also for testing is given inside the parentheses. In the last two columns the execution time of each method when 10% is used for testing is shown. Execution times that correspond to prediction error results inside parentheses are omitted. It can be seen that the

proposed method has achieved the lowest prediction error for Set \mathcal{D} , i.e. $42.0 \pm 0.3\%$. Method 7 follows with $42.7 \pm 0.3\%$. From the comparison with the prediction error achieved by a single Gaussian model, it is inferred that the proposed method improves prediction error by 6.5%. As regards Set \mathcal{E} , method 7 achieved about the same prediction error with the proposed method, i.e. about 47.4%. However, method 7 was three times faster than the proposed one can be seen from the last column. From the results inside parentheses, it is seen that methods 4, 7, and the proposed one achieved 36.9%, 39.9%, and 39.8% for Set \mathcal{D} , whereas the single Gaussian model achieves 47.5%. As regards Set \mathcal{E} , only method 7 and the proposed one improved the 48.9% achieved by a single Gaussian component modeling.

TABLE III
COMPARISON WITH OTHER EM VARIANTS FOR REAL DATA

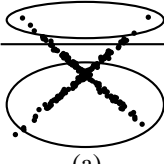
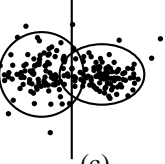
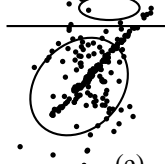
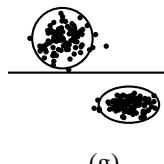
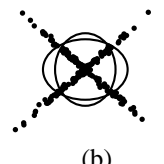
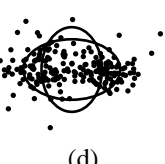
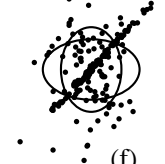
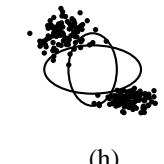
#	Method/Data-set	Prediction error (%)		Average execution time (sec)	
		\mathcal{D}	\mathcal{E}	\mathcal{D}	\mathcal{E}
1	Forward MDL-EM-Random	43.7 ± 0.3 , (41.6)	49.8 ± 0.2 , (49.8)	1.05 ± 0.03 ,	0.97 ± 0.03
2	Forward AIC-EM-Random	44.2 ± 0.3 , (40.9)	49.6 ± 0.2 , (49.5)	1.83 ± 0.05 ,	1.10 ± 0.05
3	Forward ICL-EM-Random [8]	44.0 ± 0.3 , (41.4)	49.9 ± 0.2 , (49.7)	1.26 ± 0.03 ,	1.18 ± 0.05
4	Backward MDL ₂ -EM-Random [9]	42.9 ± 0.2 , (36.9)	49.7 ± 0.2 , (48.8)	2.93 ± 0.06 ,	20.24 ± 0.29
5	Forward MDL-CEMM-Random [14]	44.2 ± 0.3 , (41.9)	49.7 ± 0.2 , (49.8)	3.12 ± 0.08 ,	3.43 ± 0.10
6	Forward MDL-DAEM-Random [13]	45.5 ± 0.3 , (43.8)	49.2 ± 0.2 , (49.3)	0.41 ± 0.01 ,	1.28 ± 0.04
7	Forward MDL-EM-Partial Random [16]	42.7 ± 0.2 , (39.9)	47.2 ± 0.2 , (46.1)	0.89 ± 0.02 ,	0.96 ± 0.02
8	Split-EM-Discriminant (Proposed)	42.0 ± 0.3 , (39.8)	47.4 ± 0.2 , (47.1)	2.66 ± 0.26 ,	3.51 ± 0.10
9	Single Gaussian modeled pdf	48.5 ± 0.2 , (47.5)	49.0 ± 0.2 , (48.9)	$0.004 \pm 4 \cdot 10^{-5}$	$0.004 \pm 6 \cdot 10^{-5}$

C. Initialization offered by the proposed MV kurtosis test

Experiments that demonstrate the advantages of the MV kurtosis test when it is used as a switch between splitting a cluster with a discriminant vs. setting the new cluster centers equal to the center of the cluster to be split are conducted. Four typical cases are shown in Table IV. The first case, shown in Figures (a) and (b) inside Table IV, depicts a clustering problem of two sources with common centers. The second case (Figures (c) and (d)) is a clustering problem when sources greatly overlap. The third case, presented in Figures (e) and (f), is another clustering problem involving two sources with common centers, which is not so symmetrical as the first case. Finally, the fourth case represents a clustering problem when sources do not overlap. The number of samples N in each case is 600 equally distributed between the sources. The MV kurtosis value for each case is shown in the first line of Table IV. From (35),

it is inferred that K_0 which is employed in the MV kurtosis test equals 7.946 for $N = 600$ and $D = 2$. In the second and the third lines of Table IV, the initialization results offered by the discriminant and the common centers methods are shown, respectively. The average correctness and the execution time of EM for 1000 Monte Carlo repetitions are included for each initialization. It is seen that the initializations that result to the highest correctness and lowest execution time are those presented in Figures (b), (c), (f), and (g). From the comparison of the MV kurtosis values of the first line with the decision threshold $K_0 = 7.946$, it is inferred that the proposed method selects the best initialization for each case.

TABLE IV
INITIALIZATION EXAMPLES

	Case 1: $K = 13.789$	Case 2: $K = 7.452$	Case 3: $K = 10.104$	Case 4: $K = 6.029$
Discriminant ($K_0 < 7.946$)	 <p>(a)</p> <p>Corr.: 30.0%, Time: 0.485</p>	 <p>(c)</p> <p>Corr.: 99.0%, Time: 0.343</p>	 <p>(e)</p> <p>Corr.: 60.8%, Time: 0.426</p>	 <p>(g)</p> <p>Corr.: 100.0%, Time: 0.273</p>
Common centers ($K_0 > 7.946$)	 <p>(b)</p> <p>Corr.: 99.2%, Time: 0.472</p>	 <p>(d)</p> <p>Corr.: 93.2%, Time: 0.417</p>	 <p>(f)</p> <p>Corr.: 98.2%, Time: 0.350</p>	 <p>(h)</p> <p>Corr.: 98.9%, Time: 0.309</p>

Abbreviations: Corr. stands for averaged correctness, and Time stands for averaged execution time, both estimated for 1000 Monte Carlo repetitions of the experiment.

VI. CONCLUSIONS

An algorithm based on expectation-maximization algorithm for clustering sample measurement vectors for any dimension has been proposed. The basic idea behind the algorithm is to employ multivariate statistical tests as plug-in criteria for splitting non-Gaussian distributed clusters to Gaussian distributed ones.

From the experiments, it is inferred that the proposed method as well as methods 4 [9] and 7 [16], are the most accurate ones. Method 7, however, sometimes fails to initialize correctly the GMM because, the partial random initialization is accomplished by keeping the old components of the GMM fixed, while refining the new component. Method 4 is found rather slow, because it assumes initially a great number of components in the mixture. The proposed method has been found to suffer sometimes from over-splitting.

This problem may be solved by changing the calculation of the confidence limits of $\hat{F}_{R_i}(r_i)$ in Section III, or by employing also the angle information between a sample measurement vector and the center of a component. The information related to the angle of a sample measurement vector from the component center is lost either in the multivariate normality test that is based on the Mahalanobis distance of each sample measurement vector from the component center or the multivariate kurtosis which is simply the sum of squares of the aforementioned Mahalanobis distances. Therefore, the proposed method can be extended with statistical tests based on the angle information to assess multivariate normality [28], [34].

APPENDIX

The assumption that Mahalanobis distance can be treated as a r.v. R_i that follows a beta distribution was extensively used in Sections II, III, and IV. The proof of this assumption is rather complex, so it will be revised here.

Let us assume that $\underline{x} = [X_1, X_2, \dots, X_D]^T$ is a D-dimensional vector that is distributed according to the multivariate (MV) normal distribution $MVN_D(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with probability density function (pdf)

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-D/2} (|\boldsymbol{\Sigma}|)^{-0.5} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \quad (44)$$

where $|\boldsymbol{\Sigma}|$ is the determinant of non-singular positive semi-definite covariance matrix $\boldsymbol{\Sigma}$. The Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}$ is defined as

$$r = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}). \quad (45)$$

In most cases, the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are unknown, and therefore, they are replaced by the sample mean vector $\bar{\mathbf{x}}$ and the sample dispersion matrix \mathbf{S} of a set of sample measurement vectors $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^N$ defined as

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{j=1}^N \mathbf{x}_j, \quad (46)$$

$$\mathbf{S} = \frac{1}{N-1} \sum_{j=1}^N (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^T. \quad (47)$$

Accordingly (45) becomes

$$r_i = (\mathbf{x}_i - \bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}}). \quad (48)$$

Let R_i be r.v.s that admit values r_i given by (48) for $i = 1, 2, \dots, N$. The distribution of Mahalanobis distance is the distribution of the r.v. R_i . In this Appendix, we will revise the following proof which is attributed to S. S. Wilks [26]. If \mathbf{x} is distributed as in (44), then R_i obeys

$$\frac{N}{(N-1)^2} R_i \sim f_{Beta}\left(\frac{N}{(N-1)^2} r_i \mid \frac{D}{2}, \frac{N-D-1}{2}\right), \quad (49)$$

where $f_{Beta}(x | a, b)$ is the beta distribution with parameters a and b , and $D < N$. The cumulative distribution function (cdf) of R_i is necessary for testing MV normality hypothesis in Section III. $F_{R_i}(r_i)$ according to (49) is

$$F_{R_i}(r_i) = I_{\frac{Nr_i}{(N-1)^2}} \left(\frac{D}{2}, \frac{N-D-1}{2} \right), \quad (50)$$

where $I_x(a, b)$ is the incomplete beta function.

The logical sequence of the proof is summarized in Figure 14. The proof that $\frac{N}{(N-1)^2} R_i$ is distributed

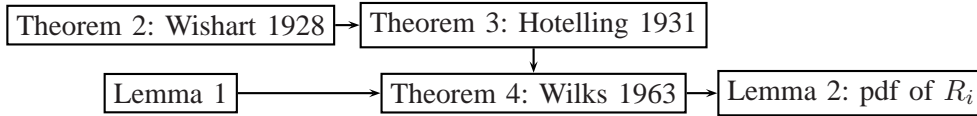


Fig. 14. Logical sequence of steps to arrive at the distribution of Mahalanobis distance.

as $f_{Beta}(\frac{N}{(N-1)^2} r_i | \frac{D}{2}, \frac{N-D-1}{2})$ is given in Lemma 2. However, before dealing with Lemma 2, first some additional theorems and lemmata should be proven. Theorem 2 defines the distribution of the sample dispersion matrix of a multivariate Gaussian R.V..

Theorem 2: The matrix \mathbf{S} follows the Wishart distribution $W_D(\mathbf{\Sigma}, N)$ with scale matrix $\mathbf{\Sigma}$ and degrees of freedom N . The pdf of $\mathbf{A} = (N-1)\mathbf{S}$ is

$$f(\mathbf{A}) = \frac{\|\mathbf{A}\|^{\frac{N-D-2}{2}} \exp\left(-\frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{A})\right)}{2^{\frac{(N-1)D}{2}} \pi^{\frac{D(D-1)}{4}} \|\mathbf{\Sigma}\|^{\frac{(N-1)}{2}} \prod_{i=1}^D \Gamma\left(\frac{N-i}{2}\right)} \quad (51)$$

where $\Gamma(\cdot)$ is the Gamma function.

Proof: See [35]. ■

The next theorem defines the distribution of the scaled Euclidean distance where relationships across dimensions are taken into account.

Theorem 3: If $T^2 = \mathbf{Y}^T \mathbf{S}^{-1} \mathbf{Y}$ where \mathbf{Y} and \mathbf{S} are independent and distributed according to $MVN_D(\mathbf{0}, \mathbf{\Sigma})$ and $W_D(\mathbf{\Sigma}, N)$ respectively, then T^2 obeys the Hotelling distribution:

$$f_{T^2}(t^2) = \frac{\Gamma(\frac{N}{2})}{(N-1)\Gamma(\frac{N-D}{2})\Gamma(\frac{D}{2})} \left(\frac{t^2}{N-1}\right)^{\frac{D}{2}-1} \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}}. \quad (52)$$

Proof: See [36], [37]. ■

Let us prove the following lemma that will be subsequently exploited in the proof of Theorem 4.

Lemma 1: If $\sum_{i=1(\xi)}^N$ denotes the sum from $i = 1$ to N excluding ξ and

$$\mathbf{A}_{(\xi)} \triangleq \sum_{i=1(\xi)}^N (\mathbf{x}_i - \bar{\mathbf{x}}_{(\xi)})(\mathbf{x}_i - \bar{\mathbf{x}}_{(\xi)})^T, \quad \text{where } \bar{\mathbf{x}}_{(\xi)} = \frac{1}{N-1} \sum_{i=1(\xi)}^N \mathbf{x}_i, \quad (53)$$

then

$$\mathbf{A}_{(\xi)} = \mathbf{A} - \frac{N}{N-1}(\mathbf{x}_\xi - \bar{\mathbf{x}})(\mathbf{x}_\xi - \bar{\mathbf{x}})^T. \quad (54)$$

Proof: It is known that

$$\mathbf{A} = \left(\sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \right) - N \bar{\mathbf{x}} \bar{\mathbf{x}}^T = \left(\sum_{i=1(\xi)}^N \mathbf{x}_i \mathbf{x}_i^T \right) + \mathbf{x}_\xi \mathbf{x}_\xi^T - N \bar{\mathbf{x}} \bar{\mathbf{x}}^T, \quad (55)$$

$$\mathbf{A}_{(\xi)} = \left(\sum_{i=1(\xi)}^N \mathbf{x}_i \mathbf{x}_i^T \right) - (N-1) \bar{\mathbf{x}}_{(\xi)} \bar{\mathbf{x}}_{(\xi)}^T, \quad (56)$$

$$N \bar{\mathbf{x}} = (N-1) \bar{\mathbf{x}}_{(\xi)} + \mathbf{x}_\xi. \quad (57)$$

Then

$$\mathbf{A} - \mathbf{A}_{(\xi)} = \mathbf{x}_\xi \mathbf{x}_\xi^T - N \bar{\mathbf{x}} \bar{\mathbf{x}}^T + (N-1) \bar{\mathbf{x}}_{(\xi)} \bar{\mathbf{x}}_{(\xi)}^T. \quad (58)$$

By replacing $\bar{\mathbf{x}}_{(\xi)}$ with (57), (54) is obtained. ■

Theorem 4: Let

$$R_{(\xi)} \triangleq \frac{\|\mathbf{A}_{(\xi)}\|}{\|\mathbf{A}\|} \quad (59)$$

be called as one-outlier scatter ratio for sample measurement vector \mathbf{x}_ξ , i.e. it denotes how much the dispersion of the whole set differs from the same set, when \mathbf{x}_ξ is excluded. $R_{(\xi)}$ follows the beta distribution $f_{Beta}(r_{(\xi)} \mid \frac{N-D-1}{2}, \frac{D}{2})$.

Proof: If a_{jk} and $a_{jk(\xi)}$, $j, k = 1, 2, \dots, D$ are the elements of \mathbf{A} and $\mathbf{A}_{(\xi)}$ respectively, then according to Lemma 1

$$a_{jk} = a_{jk(\xi)} + \frac{N}{N-1}(x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k), \quad (60)$$

where \bar{x}_j denotes the j th element of $\bar{\mathbf{x}}$. Let us denote $\|A\| = \|a_{jk}\|$ the determinant of a matrix, then from (60)

$$\|a_{jk}\| = \|a_{jk(\xi)} + \frac{N}{N-1}(x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k)\|. \quad (61)$$

So

$$\|a_{jk}\| = \|a_{jk(\xi)}\| \left[1 + \frac{N}{N-1} \sum_{j,k=1}^D a^{jk(\xi)} (x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k) \right], \quad (62)$$

where $a^{jk(\xi)}$ is the cofactor of jk th element in $\mathbf{A}_{(\xi)}^{-1}$. Therefore,

$$R_{(\xi)} = \frac{1}{1 + \frac{N}{N-1} \sum_{j,k=1}^D a^{jk(\xi)} (x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k)} = \frac{1}{1 + \frac{N}{N-1} (\mathbf{x}_\xi - \bar{\mathbf{x}})^T \mathbf{A}_{(\xi)}^{-1} (\mathbf{x}_\xi - \bar{\mathbf{x}})}. \quad (63)$$

Since, $\mathbf{A}_{(\xi)} = (N-2)\mathbf{S}_{(\xi)} \Rightarrow \mathbf{A}_{(\xi)}^{-1} = \frac{1}{(N-2)}\mathbf{S}_{(\xi)}^{-1}$, then

$$R_{(\xi)} = \frac{1}{1 + \frac{N}{(N-1)(N-2)}(\mathbf{x}_{\xi} - \bar{\mathbf{x}})^T \mathbf{S}_{(\xi)}^{-1}(\mathbf{x}_{\xi} - \bar{\mathbf{x}})}. \quad (64)$$

Since $\mathbf{x}_{\xi} - \bar{\mathbf{x}} \sim MVN_D(\mathbf{0}, \frac{N-1}{N}\mathbf{\Sigma})$:

$$\begin{aligned} \mathbf{x}_{\xi} - \bar{\mathbf{x}}_N &= \frac{N-1}{N}\mathbf{x}_{\xi} - \frac{\mathbf{x}_1 + \dots + \mathbf{x}_{\xi-1} + \mathbf{x}_{\xi+1} + \dots + \mathbf{x}_N}{N} \sim MVN_D\left(\frac{N-1}{N}\boldsymbol{\mu}, \right. \\ &\quad \left. \left(\frac{N-1}{N}\right)^2\mathbf{\Sigma}\right) - \frac{1}{N}MVN_D\left((N-1)\boldsymbol{\mu}, (N-1)\mathbf{\Sigma}\right) = MVN_D\left(\mathbf{0}, \frac{N-1}{N}\mathbf{\Sigma}\right), \end{aligned} \quad (65)$$

by assuming that $\mathbf{d}_{\xi} = \sqrt{\frac{N}{N-1}}(\mathbf{x}_{\xi} - \bar{\mathbf{x}})$, then $\mathbf{d}_{\xi} \sim MVN_D(\mathbf{0}, \mathbf{\Sigma})$. Therefore, (64) becomes

$$R_{(\xi)} = \frac{1}{1 + \frac{1}{N-2}\mathbf{d}_{\xi}^T \mathbf{S}_{(\xi)}^{-1} \mathbf{d}_{\xi}}. \quad (66)$$

According to Theorem 3 and given that $\mathbf{d}_{\xi} \sim MVN_D(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{S}_{(\xi)} \sim W_D(\mathbf{\Sigma}, N-1)$, where \mathbf{d}_{ξ} and $\mathbf{S}_{(\xi)}$ are independently distributed, because \mathbf{d}_{ξ} is not involved in the estimation of $\mathbf{S}_{(\xi)}$, it is inferred that the distribution of $T_{(\xi)}^2 = \mathbf{d}_{\xi}^T \mathbf{S}_{(\xi)}^{-1} \mathbf{d}_{\xi}$ is

$$f_{T_{(\xi)}^2}(t_{(\xi)}^2) = \frac{\Gamma(\frac{N-1}{2})}{(N-2)\Gamma(\frac{N-D-1}{2})\Gamma(\frac{D}{2})} \left(\frac{t_{(\xi)}^2}{N-2}\right)^{\frac{D}{2}-1} \left(1 + \frac{t_{(\xi)}^2}{N-2}\right)^{-\frac{N-1}{2}}. \quad (67)$$

By using the fundamental theorem for functions of one r.v. [29], the distribution of $R_{(\xi)} = \frac{1}{1 + \frac{T_{(\xi)}^2}{N-2}}$ is found as follows:

$$f_{R_{(\xi)}}(r_{(\xi)}) = \frac{f_{T_{(\xi)}^2}(t_{(\xi)}^2)}{\left|\frac{dg(t_{(\xi)}^2)}{dt^2}\right|}, \quad (68)$$

where

$$g(t_{(\xi)}^2) = \frac{1}{1 + \frac{t_{(\xi)}^2}{N-2}}, \quad (69)$$

$$\frac{dg(t_{(\xi)}^2)}{dt_{(\xi)}^2} = \left(1 + \frac{t_{(\xi)}^2}{N-2}\right)^{-2} \frac{1}{N-2}, \quad (70)$$

$$t_{(\xi)}^2 = (N-2)\left(\frac{1}{r_{(\xi)}} - 1\right). \quad (71)$$

So,

$$f_{R_{(\xi)}}(r_{(\xi)}) = \left(1 + \frac{t_{(\xi)}^2}{N-2}\right)^2 \frac{(N-2)\Gamma(\frac{N-1}{2})}{(N-2)\Gamma(\frac{N-D-1}{2})\Gamma(\frac{D}{2})} \left(\frac{1}{r_{(\xi)}} - 1\right)^{\frac{D}{2}-1} \left(\frac{1}{r_{(\xi)}}\right)^{-\frac{N-1}{2}} = \quad (72)$$

$$\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-D-1}{2})\Gamma(\frac{D}{2})} (1 - r_{(\xi)})^{\frac{D}{2}-1} r_{(\xi)}^{\frac{N-1}{2}-2+1-\frac{D}{2}}, \quad (73)$$

which is the $f_{Beta}(r_{(\xi)} | \frac{N-D-1}{2}, \frac{D}{2})$ distribution. ■

Lemma 2: If $R_{(\xi)} \sim f_{Beta}(r_{(\xi)} \mid \frac{N-D-1}{2}, \frac{D}{2})$ then

$$\frac{N}{(N-1)^2} R_i \sim f_{Beta}\left(\frac{N}{(N-1)^2} r_i \mid \frac{D}{2}, \frac{N-D-1}{2}\right). \quad (74)$$

Proof: From (60), we obtain $a_{jk(\xi)} = a_{jk} - \frac{N}{N-1}(x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k)$. Then

$$\frac{||a_{jk(\xi)}||}{||a_{jk}||} = 1 - \frac{N}{N-1} \sum_{i,j=1}^D a^{jk}(x_{\xi j} - \bar{x}_j)(x_{\xi k} - \bar{x}_k), \quad (75)$$

where a^{jk} is the cofactor of jk th element in \mathbf{A}^{-1} . Hence,

$$R_{(\xi)} = 1 - \frac{N}{N-1} (\mathbf{x}_{\xi} - \bar{\mathbf{x}})^T \mathbf{A}^{-1} (\mathbf{x}_{\xi} - \bar{\mathbf{x}}). \quad (76)$$

Given that $\mathbf{A}^{-1} = \frac{1}{N-1} \mathbf{S}^{-1}$, it is inferred that

$$R_{(\xi)} = 1 - \frac{N}{(N-1)^2} (\mathbf{x}_{\xi} - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_{\xi} - \bar{\mathbf{x}}) = 1 - \frac{N}{(N-1)^2} R_{\xi} \Rightarrow \frac{N}{(N-1)^2} R_{\xi} = 1 - R_{(\xi)}. \quad (77)$$

Since $R_{(\xi)} \sim f_{Beta}(r_{(\xi)} \mid \frac{N-D-1}{2}, \frac{D}{2}) \Rightarrow 1 - R_{(\xi)} \sim f_{Beta}(1 - r_{(\xi)} \mid \frac{D}{2}, \frac{N-D-1}{2})$ [30]. Therefore from (77), it is deduced that

$$\frac{N}{(N-1)^2} R_{\xi} \sim f_{Beta}\left(\frac{N}{(N-1)^2} r_{\xi} \mid \frac{D}{2}, \frac{N-D-1}{2}\right). \quad (78)$$

By replacing ξ with i , the proof is concluded. The result (78) is valid for every value of N and D with $D < N < \infty$. ■

REFERENCES

- [1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. (B)*, vol. 39, no. 1, pp. 1–38, 1977.
- [2] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. N.Y. Wiley, 1997.
- [3] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.
- [4] M. Law, M. Figueiredo, and A. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 26, no. 9, pp. 1154–1166, 2004.
- [5] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press, 1999.
- [6] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. 19, no. 6, p. 716, 1974.
- [7] J. Rissanen, "Stochastic complexity," *J. Roy. Stat. Soc. (B)*, vol. 49, pp. 223–239 and 253–265, 1987.
- [8] C. Biernacki, G. Celeux, and G. Govaert, "Assesing a mixture model for clustering with the integrated completed likelihood," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 22, no. 7, pp. 719–725, 2000.
- [9] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 24, no. 3, pp. 381–396, 2002.
- [10] G. Celeux and G. Soromenho, "An entropy criterion for assessing the number of clusters on a mixture model," *J. Classification*, vol. 13, pp. 195–212, 1996.
- [11] B. Zhang, C. Zhang, and X. Yi, "Competitive EM algorithm for finite mixture models," *Pat. Recognition*, vol. 37, pp. 131–144, 2004.

- [12] N. Ueda and R. Nakano, "EM algorithm with split and merge operations for mixture models," *Systems and Computers in Japan*, vol. 31, no. 5, pp. 930–940, 2000.
- [13] —, "Deterministic annealing EM algorithm," *Neural Networks*, no. 11, pp. 271–282, 1998.
- [14] G. Celeux, S. Cretien, F. Forbes, and A. Mkhadri, "A component-wise EM algorithm for mixtures," *J. Computational and Graphical Statistics*, vol. 10, pp. 669–712, 2001.
- [15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5-th Berkeley Symp. Mathematical Statistics and Probability*. Berkeley, Univ. of California Press, 1967, pp. 281–297.
- [16] J. Verbeek, M. Vlassis, and B. Kröse, "Efficient greedy learning of Gaussian mixture models," *Neural Computation*, vol. 5, no. 2, pp. 469–485, 2003.
- [17] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics & Computing*, vol. 9, p. 63, 2000.
- [18] D. Ververidis and C. Kotropoulos, "Emotional speech classification using Gaussian mixture models and the sequential floating forward selection algorithm," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2005, pp. 1500–1503.
- [19] C. Biernacki, G. Celeux, and G. Govaert, "An improvement of the NEC criterion for assessing the number of clusters in a mixture model," *Pat. Rec. Letters*, vol. 20, pp. 267–272, 1999.
- [20] G. Almpantidis and C. Kotropoulos, "Phonemic segmentation using the generalized Gamma distribution and small-sample Bayesian information criterion," *Speech Communication*, vol. 50, no. 1, pp. 38–55, 2008.
- [21] N. Vlassis and A. Likas, "A kurtosis-based dynamic approach to Gaussian mixture modeling," *IEEE Trans. Syst., Man, Cybern. A*, vol. 29, no. 4, pp. 393–399, 1999.
- [22] N. Vlassis, A. Likas, and B. Kröse, "A multivariate kurtosis-based approach to Gaussian mixture modeling," Computer Science Institute, Univ. of Amsterdam, Tech. Rep. IAS-UVA-00-04, 2000.
- [23] K. Mardia, "Measures of multivariate skewness and kurtosis with applications," *Biometrika*, vol. 57, no. 3, p. 519, 1970.
- [24] J. Koziol, "A class of invariant procedures for assessing multivariate normality," *Biometrika*, vol. 69, no. 2, p. 423, 1982.
- [25] E. Lessaffre, "Normality tests and transformations," *Pat. Rec. Letters*, vol. 1, pp. 187–199, 1983.
- [26] S. S. Wilks, *Mathematical Statistics*. N.Y. Wiley, 1962.
- [27] G. McLachlan, "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Applied Stat.*, vol. 36, no. 3, pp. 318–324, 1987.
- [28] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis*. London: Academic Press, 1979.
- [29] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. N.Y.: McGraw-Hill, 2002.
- [30] M. Evans, N. Hastings, and J. Peacock, *Statistical Distributions*. N.Y. Wiley, 2000.
- [31] R. Darlington, "Is kurtosis really peakedness," *American Statistician*, vol. 24, no. 2, pp. 19–22, 1970.
- [32] J. H. L. Hansen and B. D. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 4, pp. 307–313, 1996.
- [33] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. N.Y.: Academic Press, 1990.
- [34] J. Koziol, "On assessing multivariate normality," *J. Royal Stat. Soc.*, vol. 45, no. 3, pp. 358–361, 1983.
- [35] J. Wishart, "The generalized product moment distribution in samples from a normal multivariate population," *Biometrika*, vol. 20A, no. 1/2, pp. 32–52, 1928.
- [36] H. Hotelling, "The generalization of Student's ratio," *Annals of Math. Statistics*, vol. 2, no. 3, pp. 360–378, 1931.
- [37] T. Anderson, *An Introduction to Multivariate Statistical Analysis*. J. Wiley & Sons: N.Y., 1984.