Model Selection Criterion for Multivariate Bounded Asymmetric Gaussian Mixture Model

Zixiang Xian[1], Muhammad Azam[1], Manar Amayri[1,2], Nizar Bouguila[1]
[1]*Concordia Institute for Information Systems Engineering Concordia University, Montreal, Canada*
[2]*Grenoble Institute of Technology, Grenoble, France*
zi_xian@encs.concordia.ca, mu_azam@encs.concordia.ca, manar.amayri@grenoble-inp.fr, nizar.bouguila@concordia.ca

***Abstract*—In this paper, model selection criterion for bounded support asymmetric Gaussian mixture model (BAGMM) using minimum message length (MML) is proposed. The proposed approach is validated using synthetic data, real data and occupancy detection application. The proposed approach is compared with other state of the art model selection approaches. Moreover, the developed bounded mixture is compared with asymmetric Gaussian mixture model (AGMM).**

***Index Terms*—Multivariate Bounded Asymmetric Gaussian Mixture Model (BAGMM), Minimum Message Length (MML), Model Selection, Data Clustering, Occupancy Detection**

I. Introduction

Finite mixture models are widely used in different applications for pattern recognition, statistical inference, data mining and information retrieval. In particular, Gaussian mixture model (GMM) is a well-known approach widely used in many applications. Expectation Maximization (EM) algorithm is utilized to estimate the parameters of mixture model by maximizing the log-likelihood function efficiently [1]–[3]. However, Gaussian distribution is symmetric in nature and sensitive to outliers. Because in real life, data can be asymmetric, and in order to improve the robustness of clustering and enhance the modeling capabilities for asymmetric datasets, the asymmetric Gaussian mixture model (AGMM) has been proposed in [4]. On the other hand, the mixture of generalized Gaussian distribution (GGD) was proposed to overcome the drawback of GMM's rigidity of its shape [5], [6] and has been applied to many real applications [7].

In many real applications, data lie in a bounded support range, whereas algorithms to model these datasets have an unbounded support range. The idea of bounded support mixtures was proposed due to bounded nature of data in many applications and bounded support Gaussian mixture model (BGMM) was developed in [8]–[10] to better model real-world data. Several bounded support mixtures have been proposed so far to improve clustering [11], [12]. Bounded asymmetric Gaussian mixture model (BAGMM) has been proposed in [13] and successfully applied to several applications.

Unfortunately the research work in [13] does no involve an automatic approach to determine the optimal number of clusters. In general, there are many ways to achieve this by either deterministic or stochastic way. The general stochastic way uses Markov Chain Monte Carlo (MCMC) methods to either implement the model selection criteria [14] or approximate the posterior distribution to find the optimal clusters. In this paper, we want to focus on deterministic approaches for which several techniques have been proposed including Akaike's information criterion (AIC) [15], the Schwarz's Bayesian information criterion (BIC) [16], Consistent AIC (CAIC) [17], minimum description length (MDL) [18], the mixture minimum description length (MMDL) [19], the Laplace empirical criterion (LEC) [20] and minimum message length (MML) [4], [5]. Model selection using MML has outperformed the AIC and MDL criteria in several studies [21], [22].

This paper proposes model selection using MML for the BAGMM and compares it with other model selection criteria. The experiments are conducted on several synthetic and real datasets, including an application to occupancy detection. The clustering performance is compared with AGMM after determining the optimal number of clusters.

The rest of the paper is organized as follows: Section II presents the bounded asymmetric Gaussian mixture briefly. The proposed model selection criterion using MML is described in the Section III, including complete model learning. The experiments and results are presented in in Section IV and Section V is dedicated to conclusions and future perspectives.

II. Proposed Model

A. Mixture of Bounded Asymmetric Gaussian Distributions

Given a D-dimensional random variable $\vec{X} = (X_1, ..., X_D)$, that follows K components mixture distribution, then:

$$p(\vec{X}|\Theta) = \sum_{j=1}^{K} p(\vec{X}|\xi_j)p_j \tag{1}$$

provided $p_j \geq 0$, $\sum_{j=1}^{K} p_j = 1$, $\Theta = (\xi_1, \xi_2, \xi_3, \xi_4)$ with $\xi_1 = (\vec{\mu}_1, ..., \vec{\mu}_K)$, $\xi_2 = (\vec{\sigma}_{l_1}, ..., \vec{\sigma}_{l_K})$, $\xi_3 = (\vec{\sigma}_{r_1}, ..., \vec{\sigma}_{r_K})$ and $\xi_4 = (p_1, ..., p_K)$. The term $p(\vec{X}|\xi_j)$ is the PDF of the bounded asymmetric Gaussian distribution (BAGD) for vector $\vec{X}$ and defined as:

$$p(\vec{X}|\xi_j) = \frac{f(\vec{X}|\xi_j)\mathrm{H}(\vec{X}|\Omega_j)}{\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j)d\mathrm{u}}, \text{where } \mathrm{H}(\vec{X}|\Omega_j) = \begin{cases} 1 & \text{if } \vec{X} \in \partial_j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$f(\vec{X}|\xi_j) = \prod_{d=1}^{D} \frac{2}{\sqrt{2\pi}(\sigma_{l_{jd}} + \sigma_{r_{jd}})} \times \begin{cases} \exp\left[-\frac{(\mathrm{X}_d - \mu_{jd})^2}{2\sigma_{l_{jd}}^2}\right] & \text{if } \mathrm{X}_d < \mu_{jd} \\ \exp\left[-\frac{(\mathrm{X}_d - \mu_{jd})^2}{2\sigma_{r_{jd}}^2}\right] & \text{if } \mathrm{X}_d \geq \mu_{jd} \end{cases} \tag{3}$$

where $\vec{\mu}_j = (\mu_{j1}, ..., \mu_{jD})$, $\vec{\sigma_{l_j}} = (\sigma_{l_{j1}}, ..., \sigma_{l_{jD}})$, and $\vec{\sigma_{r_j}} = (\sigma_{r_{j1}}, ..., \sigma_{r_{jD}})$ are the mean, left standard deviation and right standard deviation of the D-dimensional BAGD, respectively.

The term $\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j)d\mathrm{u}$ in Eq. (2) is the normalization constant that indicates the share of $f(\vec{X}|\xi_j)$ which belongs to the support region ∂. We introduce stochastic indicator vectors $\vec{Z}_i = (Z_{i1}, ..., Z_{iK})$, one for each observation. The role is to encode the membership of each observation for a relative component of the mixture model. In other words, Z_{ij}, the hidden variable in each indicator vector, equals 1 if $\vec{X}_i$ belongs to class j and 0, otherwise. The complete data likelihood is given below:

$$p(\mathcal{X}, \mathcal{Z}|\Theta) = \prod_{i=1}^{N}\prod_{j=1}^{K}\Big(p(\vec{X}_i|\xi_j)p_j\Big)^{Z_{ij}} \tag{4}$$

where Z_{ij} is the posterior probability and can be written as:

$$Z_{ij} = p(j|\vec{X}_i) = \frac{p(\vec{X}_i|\xi_j)p_j}{\sum_{j=1}^{K} p(\vec{X}_i|\xi_j)p_j} \quad \text{and} \quad \mathcal{Z} = \{\vec{Z}_1, ..., \vec{Z}_N\}. \tag{5}$$

The parameters are estimated using EM algorithm, which is adopted from [13].

III. Model selection using minimum message length (MML) criterion

The general form of MML equation, which we should minimize to obtain the optimal number of clusters in the mixture, is as follows:

$$\text{Mess Len}(K) \simeq -\log(p(\Theta_K)) - \mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{1}{2}\log|F(\Theta_K)| + \frac{N_p}{2}\left(1 + \log\left(\frac{1}{12}\right)\right) \tag{6}$$

where N_p is number of parameters (equal to $K(3D+1)$), Θ_K is set of parameters when mixture contains K components, $p(\Theta_K)$ is prior probability, $\mathcal{L}(\Theta_K, Z, \mathcal{X})$ is log-likelihood of mixture model and $|F(\Theta_K)|$ is determinant of Fisher information matrix. The estimation of number of classes is carried out by finding minimum with respect to Θ of message length [4], [5]. The derivation of $p(\Theta_K)$ and $|F(\Theta_K)|$ is given in following subsections.

A. Derivation of the prior $p(\Theta)$

we assume that all the parameters of the mixture model are mutually independent, then the prior distribution over the parameters, π, μ, σ_l and σ_r, is :

$$p(\Theta) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}_l)p(\boldsymbol{\sigma}_r) \tag{7}$$

where $\boldsymbol{\pi} = (p_1, ..., p_K)$. Each parameter is independent, so each prior distribution is defined separately. Beginning with $p(\boldsymbol{\pi})$, we know that vector $\boldsymbol{\pi}$ is defined on the simplex as $\{(p_1, ..., p_K) : \sum_{j=1}^{K} p_j = 1\}$. In general, the Dirichlet distribution is a natural choice as a prior for vector $\boldsymbol{\pi}$, which is defined as:

$$p(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{j=1}^{K}\eta_j)}{\sum_{j=1}^{K}\Gamma(\eta_j)}\sum_{j=1}^{K} {p_j}^{\eta_j - 1} \tag{8}$$

where $(\eta_1, ..., \eta_K)$ is the parameters vector of Dirichlet distribution. By choosing, $\eta_1 = 1, ..., \eta_K = 1$, we get a uniform prior over space $p_1 + ... + p_K = 1$, which is represented as: $p(\boldsymbol{\pi}) = (K-1)!$. For each μ_{jd}, uniform prior is chosen. Each μ_{jd} is chosen to be uniform in the region $(\mu_{jd} - \sigma_{l_d} \leq \mu_{jd} \leq \mu_{jd} + \sigma_{r_d})$, then prior for μ_j is given by the following equations:

$$p(\mu_{jd}) = \frac{1}{\sigma_{ld} + \sigma_{rd}} \implies p(\vec{\mu}_j) = \prod_{d=1}^{D}\frac{1}{\sigma_{ld} + \sigma_{rd}} \tag{9}$$

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{K}\prod_{d=1}^{D}\frac{1}{\sigma_{ld} + \sigma_{rd}} = \prod_{d=1}^{D}\frac{1}{(\sigma_{ld} + \sigma_{rd})^K} \tag{10}$$

For the parameter σ_l and σ_r, we have:

$$p(\boldsymbol{\sigma}_l) = \prod_{j=1}^{K} p(\vec{\sigma}_{l_j}), \quad p(\boldsymbol{\sigma}_r) = \prod_{j=1}^{K} p(\vec{\sigma}_{r_j}) \tag{11}$$

where different components of vectors $\vec{\sigma}_{l_j}$ and $\vec{\sigma}_{r_j}$ are assumed to be independent. The principle of ignorance is adopted due to the absence of other knowledge about $\sigma_{l_{jd}}$ and $\sigma_{r_{jd}}$, by taking from a uniform prior. The $\vec{\mu} = (\mu_1, ..., \mu_D)$, $\vec{\sigma_l} = (\sigma_{l_1}, ..., \sigma_{l_D})$ and $\vec{\sigma_r} = (\sigma_{r_1}, ..., \sigma_{r_D})$ are mean, left standard deviation and right standard deviation vectors of whole dataset, respectively. And for each $\sigma_{l_{jd}}$ and $\sigma_{r_{jd}}$, following uniform prior will be used:

$$p(\sigma_{l_{jd}}) = \frac{1}{\sigma_{l_d}}, \quad p(\sigma_{r_{jd}}) = \frac{1}{\sigma_{r_d}} \tag{12}$$

where $0 \leq \sigma_{l_{jd}} \leq \sigma_{l_d}$ and $0 \leq \sigma_{r_{jd}} \leq \sigma_{r_d}$. It follows that

$$p(\vec{\sigma}_{l_j}) = \prod_{d=1}^{D}\frac{1}{\sigma_{l_d}}, \quad p(\vec{\sigma}_{r_j}) = \prod_{d=1}^{D}\frac{1}{\sigma_{r_d}} \tag{13}$$

From Eqs. (11 & 13), we obtain:

$$p(\boldsymbol{\sigma}_l) = \prod_{j=1}^{K}\prod_{d=1}^{D}\frac{1}{\sigma_{l_d}} = \prod_{d=1}^{D}\frac{1}{{\sigma_{l_d}}^K}, \quad p(\boldsymbol{\sigma}_r) = \prod_{j=1}^{K}\prod_{d=1}^{D}\frac{1}{\sigma_{r_d}} = \prod_{d=1}^{D}\frac{1}{{\sigma_{r_d}}^K} \tag{14}$$

Finally, by replacing the priors of parameters in Eq. (7) by Eqs. (10 & 14), we get:

$$p(\Theta) = (M-1)!\prod_{d=1}^{D}\frac{1}{{\sigma_{l_d}}^K {\sigma_{r_d}}^K (\sigma_{ld} + \sigma_{rd})^K} \tag{15}$$

B. Derivation of the Fisher information matrix $|F(\Theta)|$

In general, the Fisher information matrix is the expected value of the Hessian matrix minus the log-likelihood. But in practice, it is intractable to compute the expected Fisher information matrix. So we utilize the complete-data Fisher information matrix to approximate the Hessian matrix, which is the product of the information matrix's determinant for each cluster times the information matrix of the mixing weight as in Eq. (16).

$$|F(\Theta)| = |F(\boldsymbol{\pi})||F(\boldsymbol{\mu})||F(\boldsymbol{\sigma}_l)||F(\boldsymbol{\sigma}_r)| \tag{16}$$

$$|F(\boldsymbol{\pi})| = \frac{N^{K-1}}{\sum_{j=1}^{K} p_j}, \quad F(\vec{\mu}_j)_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial\mu_{jk_1}\partial\mu_{jk_2}} \tag{17}$$

$$F(\vec{\sigma}_{l_j})_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial\sigma_{l_{jk_1}}\partial\sigma_{l_{jk_2}}}, \quad F(\vec{\sigma}_{r_j})_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial\sigma_{r_{jk_1}}\partial\sigma_{r_{jk_2}}} \tag{18}$$

$$\begin{aligned}
|F(\boldsymbol{\mu})| = \prod_{j}^{K}\prod_{d=1}^{D}\Bigg| & \sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\left[\frac{-1}{\sigma_{l_{jd}}^2}\right] + \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\left[\frac{-1}{\sigma_{r_{jd}}^2}\right] \\
& - \sum_{i=l,\mathrm{x}_d<\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{l_{jd}}^4}\times\Big\{-\frac{\left(\frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}m_{jd}-\mu_{jd})\mathrm{H}(\mathrm{l}m_{jd}|\Omega_j)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}m_{jd}|\Omega_j)\right)^2} \\
& + \frac{\frac{1}{M}\sum_{m=1}^{M}\left\{(\mathrm{l}m_{jd}-\mu_{jd})^2 - 1\right\}\mathrm{H}(\mathrm{l}m_{jd}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}m_{jd}|\Omega_j)}\Big\} \\
& - \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{r_{jd}}^4}\times\Big\{-\frac{\left(\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}m_{jd}-\mu_{jd})\mathrm{H}(\mathrm{r}m_{jd}|\Omega_j)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}m_{jd}|\Omega_j)\right)^2} \\
& + \frac{\frac{1}{M}\sum_{m=1}^{M}\left\{(\mathrm{r}m_{jd}-\mu_{jd})^2 - 1\right\}\mathrm{H}(\mathrm{r}m_{jd}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}m_{jd}|\Omega_j)}\Big\}\Bigg|
\end{aligned} \tag{19}$$

$$|F(\boldsymbol{\sigma}_l)| = \prod_{j}^{K}\prod_{d=1}^{D}\Big|-3\sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{l_{jd}}^4}\right)$$
$$-\sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\left(\frac{-2}{\sigma_{l_{jd}}^3(\sigma_{l_{jd}}+\sigma_{r_{jd}})}\right)\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=1,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{l_{jd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}}-\mu_{jd})^4\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\frac{-3}{\sigma_{l_{jd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{l_{jd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)\right)^2}{(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j))^2}\right\}\Big| \tag{20}$$

$$|F(\boldsymbol{\sigma}_r)| = \prod_{j}^{K}\prod_{d=1}^{D}\Big|-3\sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{r_{jd}}^4}\right)$$
$$-\sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\left(\frac{-2}{\sigma_{r_{jd}}^3(\sigma_{l_{jd}}+\sigma_{r_{jd}})}\right)\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{r_{jd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^4\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\frac{-3}{\sigma_{r_{jd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\}$$
$$-\sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1}\frac{1}{\sigma_{r_{jd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)\right)^2}{(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j))^2}\right\}\Big| \tag{21}$$

where $\mathrm{l}_{m_{jd}}$ is a set of random variables drawn from the asymmetric Gaussian distribution (AGD) with the constraint, $\mathrm{u} < \mu_{jd}$ for the particular component j of the mixture model. These random variables have M vectors with D dimensions. M is a large integer chosen to generate the set of random variables, for example 2,000 draws in this paper. Similarly, $\mathrm{r}_{m_{jd}}$ is the random variables drawn from the AGD with constraint, $\mathrm{u} \geq \mu_{jd}$ for the particular component j of the mixture model.

C. Complete learning algorithm for BAGMM with MML

In this section, we summarize the model learning algorithm for the bounded AGMM and the model selection. we apply K-Means to initialize parameters, then use the EM algorithm to estimate the mixture parameters. During each iteration, we need to update bounded support range. Note that we initialize both left and right standard deviations with the standard deviation values obtained from each cluster by K-means. Finally, we need to set up the predefined threshold t_{min} for the log-likelihood between the two successive iterations, j and $j+1$, and a certain number of iterations, $epoch_{max}$. Once the log-likelihood difference is smaller than the preset point, the EM will converge, or it will stop after specific number of iterations to avoid the infinity loop, or it will stop when the parameters don't change any more.

After using the EM algorithm to learn the model parameters, we calculate the associated criterion MML using Eq. (6). Finally, select the optimal number of cluster K^* such that $K^* = \arg\min MML(K)$. The complete learning procedure for BAGMM with MML is given in Algorithm 1.

Algorithm 1 Model Learning for BAGMM

Input:Dataset $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$, t_{min}, K_{min}, K_{max}.
Output: Θ, $\mathcal{Z}$, K^*.
for $K_{min} \leq K \leq K_{max}$ do
  {Initialization}:
  K-Means (Compute $\vec{\mu}_1, \ldots, \vec{\mu}_K$ & cluster assignment)
  for all $1 \leq j \leq K$ do
    Computation of p_j and $\{(\vec{\sigma}_{l_j}$ & $\vec{\sigma}_{r_j})$=$\vec{\sigma}_j\}$
  {Expectation Maximization}:
  while relative change in log-likelihood $\geq t_{min}$ **or** iterations $\leq epoch_{max}$ **or** relative changes of parameters $\geq t_{min}$ do
    {[**E Step**]}:
    for all $1 \leq j \leq K$ do
      Compute $p(j|\vec{X}_i)$ for $i = 1, \ldots, N$.
    {[**M step**]}:
    update bounded support range
    for all $1 \leq j \leq K$ do
      Estimate p_j, $\vec{\mu}_j$, $\vec{\sigma}_{l_j}$ & $\vec{\sigma}_{r_j}$
  end while
  Compute $K^* = \arg\min MML(K)$
end for

IV. Experimental Results

A. Comparison with other model selection criteria

The model selection criteria selected to compare with MML, include MDL [18], AIC [15], Bayesian inference criterion (BIC) [16], consistent AIC (CAIC) [17], mixture MDL (MMDL) [19], MML_{like} [23], LEC [24], [25]. The details of these algorithms is given in [26].

B. Synthetic Datasets

We compared different model selection criteria when deploying BAGMM and AGMM with 2-dimensional synthetic datasets, sampled from the asymmetric Gaussian distribution having 2, 3, 4 and 5 clusters. The parameters of each cluster of synthetic dataset are given in Table I and each cluster has 2,000 data instances. The MML criterion along with EM algorithm of BAGMM is applied to determine the optimal number of clusters in each dataset. The clustering accuracy is also determined after finding the correct number of mixture components and results are compared with other model selection criteria and AGMM. The comparison between the AGMM and BAGMM for all the model selection criteria is provided in Table I, which demonstrates that all model selection criteria including MML for BAGMM have correctly determined the number of clusters. However, model selections criteria for AGMM, provide wrong number of clusters in each case. Table II shows the execution time and accuracy of BAGMM and AGMM under this synthetic dataset. Note that BAGMM always has high clustering accuracy as compared to AGMM, which indicates the clustering capabilty of BAGMM.

All experiments are running on a Macbook Pro 2015 with Dual-Core Intel Core i5 CPU. The BAGMM is as relatively fast as the AGMM for 5 clusters or more, but in the case of less than 5 clusters, the AGMM is a little bit faster. The BAGMM always converges faster than the AGMM with less iterations.

C. Real Datasets

We have adopted 10 standard multidimensional datasets to validate the proposed model with real datasets, which

TABLE I: The model selection and clustering results for synthetic dataset

Synthetic Dataset(2,000 instances in each cluster)									
clusters	μ, σ_l, σ_r	AIC	BIC	CAIC	MDL	MMDL	MML like	LEC	MML
2	(2, -4) , (2, 3) , (1, 5) (5, 4), (3, 6), (2.1, 3.8)	2	2	2	2	2	2	2	2
3	(2, -4), (2, 3), (1, 5) (5, 4), (3, 6), (2.1, 3.8) (-10, 12), (3, 3.7), (3.4, 5.9)	3	3	3	3	3	3	3	3
4	(2, -4), (2, 3), (1, 5) (5, 4), (3, 6), (2.1, 3.8) (-10, 12), (3, 3.7), (3.4, 5.9) (-13, 14), (1, 2.1), (3, 3)	4	4	4	4	4	4	4	4
5	(2, -4), (2, 3), (1, 5) (5, 4), (3, 6), (2.1, 3.8) (-10, 12), (3, 3.7), (3.4, 5.9) (-13, 14), (1, 2.1), (3, 3) (-15, 16.6),(3.3, 4.4), (2.8, 2.7)	5	5	5	5	5	5	5	5

TABLE II: Execution information of MML on synthetic dataset

Execution information on synthetic dataset(seconds)				
Mixture Models	Clusters	Time	Accuracy	Iterations
BAGMM	2 clusters	2.35	**71.3%**	5
BAGMM	3 clusters	8.60	**85.7%**	2
BAGMM	4 clusters	12.09	**72.2%**	4
BAGMM	5 clusters	12.58	**65.7%**	5

include Indian Liver Patient, Iris, Vertebral Column, Wine Quality (red), Spect Heart, Cryotherapy, Immunotherapy, Statlog (Heart), Parkinsons and Haberman Survival. They are from the machine learning repository at the University of California, Irvine [27]. They all differ in the number of instances, dimensions, clusters, and complexity.

The model selection using MML for BAGMM is applied on all datasets to determine the optimal number of clusters in the datasets along with comparison models and similar settings for AGMM. The description of these datasets and model selection results are presented in the Table III. It is evident from the results that MML and LEC have successfully determined the correct number of clusters in all cases with BAGMM. In the case of AGMM, MML and LEC also have a high probability of determining the correct number of clusters, however the performance with BAGMM is more accurate. The equation of MML is almost the same as the LEC, containing both prior distribution and the Fisher information matrix, which outperforms other model selection criteria.

D. Occupancy Detection and Model Selection

Occupancy detection is widely used in smart buildings and it helps in energy efficiency, improves thermal comfort and reduces carbon footprints. This section compares several model selection methods with MML using AGMM and BAGMM on an occupancy dataset. The dataset [28] is composed of 9752 instances, 5 dimensions and 2 clusters, as shown in the Table IV. In this application, we need to detect room occupancy as a binary classification from CO2, light, Humidity, temperature, and humidity ratio, which were taken every minute. Compared

TABLE III: The model selection results for real dataset

Real Dataset												
dataset	N	D	K	AIC	BIC	CAIC	MDL	MMDL	MML like	LEC	MML
Indian Liver Patient(AGMM)	583	10	**2**	4	2	2	2	4	4	2	2
Indian Liver Patient(BAGMM)				2	2	2	2	2	2	**2**	**2**
Iris(AGMM)	150	4	**3**	6	3	3	3	3	6	6	6
Iris(BAGMM)				6	6	6	6	6	6	**3**	**3**
Vertebral(AGMM)	310	6	**3**	3	3	3	3	3	3	3	3
Vertebral(BAGMM)				5	3	3	3	5	5	**3**	**3**
Wine Quality(red)(AGMM)	1599	11	**6**	5	5	5	5	5	5	6	6
Wine Quality(red)(BAGMM)				8	8	8	8	8	8	**6**	**6**
Spect Heart(AGMM)	80	44	**2**	6	4	2	4	4	6	2	2
Spect Heart(BAGMM)				5	2	2	2	5	5	**2**	**2**
Cryotherapy(AGMM)	90	6	**2**	2	2	2	2	2	2	2	2
Cryotherapy(BAGMM)				6	2	2	2	6	6	**2**	**2**
Immunotherapy(AGMM)	90	7	**2**	3	2	2	2	3	3	2	2
Immunotherapy(BAGMM)				2	2	2	2	2	2	**2**	**2**
Statlog(Heart)(AGMM)	270	13	**2**	6	6	2	6	6	6	6	6
Statlog(Heart)(BAGMM)				2	2	2	2	2	2	**2**	**2**
Parkinsons(AGMM)	197	22	**2**	6	6	6	6	6	6	6	6
Parkinsons(BAGMM)				2	2	2	2	2	2	**2**	**2**
Haberman Survival(AGMM)	306	3	**2**	2	2	2	2	2	2	2	2
Haberman Survival(BAGMM)				2	2	2	2	2	2	**2**	**2**

with 79% accuracy in AGMM, the BAGMM has shown better performance with 94.8% accuracy, because the attributes are all environmental data with a specific bounded range. It takes the BAGMM 3.66 seconds to converge within 6 epochs, while 2.04 seconds for the AGMM with 51 iterations. Figure. 1 displays the results of different model selection criteria. The hollow black circle in each graph indicates the minimum value on the y-axis and the optimal number of clusters on the x-axis. For model selection, we can conclude BAGMM with MML and other criteria has better performance in finding the number of clusters, since all model selection methods with AGMM have shown 5 as the optimal number of clusters, while the ground truth is 2.

TABLE IV: Occupancy estimation and model selection results

Models	N	D	K	AIC	BIC	CAIC	MDL	MMDL	MML like	LEC	MML	Acc
AGMM	9752	5	**2**	5	5	5	5	5	5	5	5	79%
BAGMM				2	2	2	2	2	2	**2**	**2**	**94.8%**

V. Conclusion

This paper discusses the clustering using BAGMM and proposes the MML as model selection criterion to determine the optimal number of clusters. Bounded support mixture has demonstrated its success in many clustering applications and finding the optimal number of clusters is significant in a clustering task. The proposed model is applied to synthetic datasets, real datasets and an application is developed for occupancy detection. The results demonstrate that MML outperforms the other model selection criteria. High accuracy of 94.8% is achieved for occupancy detection and MML has successfully determined the correct number of clusters. From all the experimental results, it is observed that the BAGMM and the MML provide strong modeling ability for high-dimensional and complex datasets.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal*

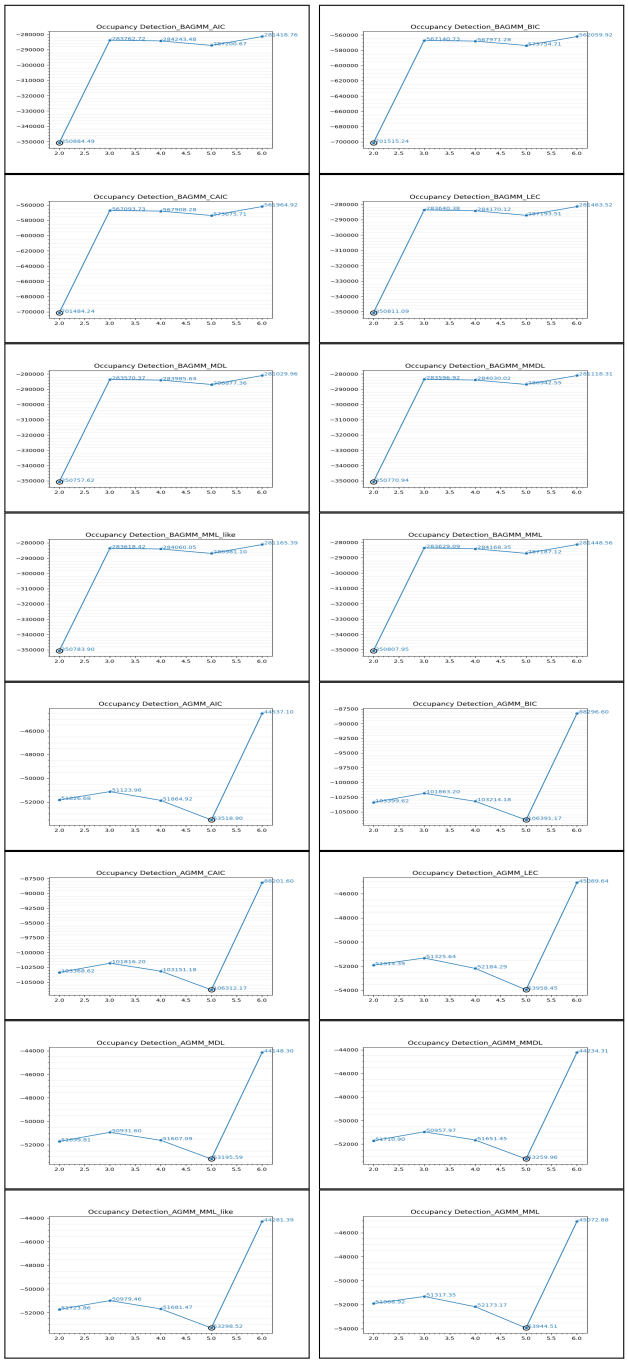

Fig. 1: Different Model Selection Criteria for Occupancy Dataset
First four rows are for **BAGMM**, Last four rows for **AGMM**

Statistical Society: Series B (Methodological), vol. 39, no. 1, pp. 1–22, 1977.

[2] F. Gu, H. Zhang, W. Wang, and S. Wang, "An expectation-maximization algorithm for blind separation of noisy mixtures using gaussian mixture model," *Circuits, Systems, and Signal Processing*, vol. 36, no. 7, pp. 2697–2726, 2017. [Online]. Available: http://dx.doi.org/10.1007/s00034-016-0424-2

[3] M.-S. Yang, C.-Y. Lai, and C.-Y. Lin, "A robust em clustering algorithm for gaussian mixture models," *Pattern Recognition*, vol. 45, no. 11, pp. 3950–3961, 2012.

[4] T. Elguebaly and N. Bouguila, "Background subtraction using finite mixtures of asymmetric Gaussian distributions and shadow detection," *Machine Vision and Applications*, vol. 25, no. 5, pp. 1145–1162, 2014.

[5] M. S. Allili, N. Bouguila, and D. Ziou, "Finite General Gaussian Mixture Modeling and Application to Image and Video Foreground Segmentation," *Journal of Electronic Imaging*, vol. 17, no. 1, pp. 013 005–013 005, 2008.

[6] M. Allili, "Wavelet Modeling Using Finite Mixtures of Generalized Gaussian Distributions: Application to Texture Discrimination and Retrieval," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1452–1464, April 2012.

[7] M. N. Do and M. Vetterli, "Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance," *IEEE transactions on image processing*, vol. 11, no. 2, pp. 146–158, 2002.

[8] P. Hedelin and J. Skoglund, "Vector quantization based on gaussian mixture models," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 4, pp. 385–401, Jul 2000.

[9] J. Lindblom and J. Samuelsson, "Bounded Support Gaussian Mixture Modeling of Speech Spectra," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 1, pp. 88–99, Jan 2003.

[10] M. Azam and N. Bouguila, "Multivariate-bounded gaussian mixture model with minimum message length criterion for model selection," *Expert Systems*, p. e12688, 2021.

[11] T. M. Nguyen, Q. J. Wu, and H. Zhang, "Bounded Generalized Gaussian Mixture Model," *Pattern Recognition*, vol. 47, no. 9, 2014.

[12] M. Azam and N. Bouguila, "Multivariate bounded support laplace mixture model," *Soft Computing*, pp. 1–30, 2020.

[13] M. Azam, B. Alghabashi, and N. Bouguila, *Multivariate Bounded Asymmetric Gaussian Mixture Model*. Cham: Springer International Publishing, 2020, pp. 61–80. [Online]. Available: https://doi.org/10.1007/978-3-030-23876-6_4

[14] N. Bouguila and D. Ziou, "A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2009.

[15] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, December 1974.

[16] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[17] H. Bozdogan, "Model selection and akaike's information criterion (aic): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[18] J. Rissanen, *Stochastic complexity in statistical inquiry*. World scientific, 1998, vol. 15.

[19] M. A. Figueiredo, J. M. Leitão, and A. K. Jain, "On fitting mixture models," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 1999, pp. 54–69.

[20] G. McLachlan and D. Peel, "Finite mixture models.,(john wiley & sons: New york.)," 2000.

[21] N. Bouguila and D. Ziou, "High-dimensional unsupervised selection and estimation of a finite generalized dirichlet mixture model based on minimum message length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1716–1731, Oct 2007.

[22] J. W. Comley and D. L. Dowe, "11 minimum message length and generalized bayesian nets with asymmetric languages," *Minimum*, p. 265, 2005.

[23] M. A. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 3, pp. 381–396, 2002.

[24] S. J. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian approaches to Gaussian mixture modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1133–1142, Nov 1998.

[25] D. Peel and G. McLachlan, "Robust Mixture Modelling using the t distribution," *Statistics and Computing*, vol. 10, no. 4, pp. 339–348, 2000.

[26] N. Bouguila, D. Ziou, and R. I. Hammoud, "A Bayesian Non-Gaussian Mixture Analysis: Application to Eye Modeling," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, June 2007, pp. 1–8.

[27] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[28] L. M. Candanedo and V. Feldheim, "Accurate occupancy detection of an office room from light, temperature, humidity and co2 measurements using statistical learning models," *Energy and Buildings*, vol. 112, pp. 28 – 39, 2016.