Statistical Modeling Using Bounded Asymmetric Gaussian Mixtures: Application to Human Action and Gender Recognition

Zixiang Xian, Muhammad Azam, Nizar Bouguila
Concordia Institute for Information Systems Engineering Concordia University, Montreal, Canada
zi_xian@encs.concordia.ca, mu_azam@encs.concordia.ca, nizar.bouguila@concordia.ca

***Abstract*—To determine the structure of high dimensional data without knowing the number of clusters nor the importance of the involved features, we propose an unsupervised feature selection framework using the bounded asymmetric Gaussian mixture model (BAGMM-FS). The bounded asymmetric Gaussian distribution has an asymmetric shape and bounded range, making it a good choice for modeling real-world data. We propose a parameter learning approach based on the expectation-maximization (EM) algorithm, and we approach the model selection task using the minimum message length (MML) criterion. The validation involves several human-related recognition challenges, such as human activity categorization and human gender recognition. It's examined from all experiments and results that BAGMM-FS has good modeling capabilities and outperforms other comparable mixture models, especially for high dimensional complex datasets.**

***Index Terms*—Bounded asymmetric Gaussian mixture model (BAGMM), Expectation-maximization (EM), Feature selection, Model selection, Minimum message length (MML), Activity categorization, Gender recognition.**

I. Introduction

Finite mixture models [1], [2] are widely applied in a wide range of machine learning applications because of their sound mathematical basis as an unsupervised learning approach. Research areas in which mixture models have been applied include data mining, image processing, computer vision, pattern recognition, etc. [2]. The most common finite mixture model, the Gaussian mixture model (GMM), has been widely studied in the past. However, the Gaussian distribution assumes that the data is symmetric and has an infinite range, which prevents it from having a good modeling capability in the presence of outliers. Therefore, some researchers have put forward the generalized Gaussian mixture model (GGMM) [3], [4], which can consider different shapes by changing its shape parameters that control the distribution's tail. Other research works have focused on the distribution support to make it more suitable for real-world data, which are always defined with bounded range. These research works include the bounded Gaussian mixture model (BGMM) that has been studied in [5], and the bounded generalized Gaussian mixture model (BGGMM), which augmented GGMM with bounded range as discussed in [6], [7]. Although the BGGMM provides higher flexibility for modeling various data shapes with bounded support, it is still symmetric that is inappropriate to model non-symmetrical data. Note that most real-world data are asymmetric which is especially true in natural images, as shown in [8]. The asymmetric Gaussian mixture model (AGMM) has been proposed to tackle that problem by having two variance parameters controlling the left and right parts of the distribution [9], [10]. Some latest research works have shown that bounded asymmetric mixture models, which we will consider in this paper, are reliable for image processing and computer vision applications [11]. Moreover, the work [12] proposed the bounded asymmetric Gaussian mixture model (BAGMM) and has shown that it generally performs better than the AGMM in clustering tasks [12].

The most general approach for parameter estimation in mixture models is based on maximizing the likelihood function through the expectation-maximization (EM) framework [13]. Generally, the EM algorithm requires an appropriate number of clusters found by model selection criteria. From a computational perspective, model selection approaches can be categorized into three classes as stochastic (e.g. Markov Chain Monte Carlo), deterministic, and re-sampling methods. Deterministic approaches that have been applied in the case of mixture models include Akaike's information criterion (AIC) [14], Schwarz's Bayesian information criterion (BIC) [15], the Laplace empirical criterion (LEC) [16] and minimum message length (MML) [9], [17], etc. More details about model selection can be found in [1]. The MML has been shown to have better performance among most model selection criteria in most cases. Apart from model selection, a crucial problem in real life applications is related to high-dimensional data. In theory, the more features we have to represent a given object, the better performance we obtain for mixture-based modeling. However, in many cases irrelevant features can compromise the effectiveness of clustering and increase the computational complexity. Hence, irrelevant feature should be given small weights or even discarded. Selecting a relevant feature space is generally known as feature selection and sometimes also called variable selection or subset selection. Although feature selection has been mainly discussed in the context of supervised learning [18], there also have been some unsupervised feature selection techniques and some of them have been proposed in the context of mixture models (see, for instance, [19]–[21]). This paper investigates the effectiveness of feature selection using AGMM (AGMM-FS) in several human related recognition challenging tasks such as activity

recognition and gender recognition [22]. The learning of the parameters is performed using MML and the resulting model is compared with other well-known mixture models using various clustering metrics.

The remainder of this paper is organized as follows: After the introduction, we present the bounded asymmetric Gaussian mixture model with feature selection (BAGMM-FS) in detail in Section II. In Section III, we develop the model's learning approach and give the complete learning algorithm. The Section IV presents the experimental results on some challenging real-world applications where the BAGMM-FS is compared with other models. The conclusion and future works are presented in Section V.

II. Bounded Asymmetric Gaussian Mixture Model

Given a D-dimensional random variable $\vec{X} = [X_1, \ldots, X_D]$ that follows a K-component mixture distribution, its probability density function (PDF) can be written as:

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

where p_j are the mixing weights that satisfy $p_j \geq 0$, $\sum_{j=1}^{K} p_j = 1$, ξ_j is the parameter of the distribution associated with jth cluster and $\Theta = (\xi_1, \cdots, \xi_K, p_1, \cdots, p_K)$ is the complete set of parameters of the AGMM. The PDF associated with each component is the multidimensional asymmetric Gaussian distribution (AGD):

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

where $\xi_j = (\vec{\mu}_j, \vec{\sigma_{l_j}}, \vec{\sigma_{r_j}})$ represents the parameters of AGD. Here, $\vec{\mu}_j = (\mu_{j1}, ..., \mu_{jD})$, $\vec{\sigma_{l_j}} = (\sigma_{l_{j1}}, ..., \sigma_{l_{jD}})$, 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 AGD, respectively. The bounded asymmetric Gaussian distribution (BAGD) for the vector $\vec{X}$ can be written as:

$$p(\vec{X}|\xi_j) = \frac{f(\vec{X}|\xi_j)\mathrm{H}(\vec{X}|\Omega_j)}{\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j)d\mathrm{u}} \tag{3}$$

$$\text{where} \quad \mathrm{H}(\vec{X}|\Omega_j) = \begin{cases} 1 & \text{if} \quad \vec{X} \in \partial_j \\ 0 & \text{otherwise} \end{cases}$$

Here, $f(\vec{X}|\xi_j)$ is the AGD, the term $\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j)d\mathrm{u}$ in Eq. (3) is the normalized constant that shows the share of $f(\vec{X}|\xi_j)$ which belongs to the support region ∂. Consider a set of independent and identically distributed vectors represented by $\mathcal{X} = (\vec{X}_1, \cdots, \vec{X}_N)$, arising from a mixture of BAGDs with K components, then its log-likelihood function can be defined as follows:

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

We introduce stochastic indicator vectors $\vec{Z}_i = (Z_{i1}, ..., Z_{iK})$, which satisfy $Z_{ij} \in \{0, 1\}$, $\sum_{j=1}^{K} Z_{ij} = 1$. In other words, Z_{ij}, the hidden variable in each indicator vector, equals 1 if $\vec{X}_i$ belongs to component j and 0, otherwise. The complete data likelihood is given by:

$$p(\mathcal{X}, \mathcal{Z}|\Theta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left(p(\vec{X}_i|\xi_j)p_j\right)^{Z_{ij}} \tag{5}$$

We can get the complete data log-likelihood by taking the logarithm of Eq. (5) as follows.

$$\log p(\mathcal{X}, Z \mid \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \log \left[p\left(\vec{X}_i \mid \xi_j\right) p_j\right] \tag{6}$$

where $\mathcal{Z} = \left\{\vec{Z}_1, \ldots, \vec{Z}_N\right\}$. According to Eq. (6), all the D features in the model have the same weight which can not describe well real-world data since some of features may be irrelevant for a some specific tasks. In order to take into account the irrelevant features, we represent them by background Gaussian distribution with parameters $\vec{\lambda} = \{\vec{\eta}, \vec{\delta}\}$, where $\vec{\eta}$ and $\vec{\delta}$ represent the mean and standard deviation, respectively. We adopt the feature relevancy approach proposed in [17] in the case of the finite Gaussian mixture. Then, the resulting model can be rewritten as:

$$p(\vec{X}_i \mid \Theta, \vec{\lambda}, \vec{\varphi}) = \sum_{j=1}^{K} p_j \prod_{d=1}^{D} p\left(X_d \mid \xi_{jd}\right)^{\varphi_d} p\left(X_d \mid \lambda_d\right)^{1-\varphi_d} \tag{7}$$

where $\vec{\varphi} = (\varphi_1, \cdots, \varphi_d)$ is a set of binary parameters such that if $\varphi_d = 1$ then dth feature is relevant, otherwise, $\varphi_d = 0$ for irrelevant features. Here, $\vec{\varphi}$ is considered as a hidden variable, and according to [17], we can obtain

$$p\left(\vec{X}_i \mid \Theta_K\right) = \sum_{j=1}^{K} p_j \prod_{d=1}^{D} \left[\omega_d p\left(X_d \mid \xi_{jd}\right) + (1 - \omega_d) p\left(X_d \mid \lambda_d\right)\right] \tag{8}$$

From above equation, we assume that not all the feature have the same relevancy by assigning weights to these features, denoted as $\vec{\omega} = (\omega_1, \cdots, \omega_D)$, where $0 \leq \omega_d \leq 1$, $d = 1, \ldots, D$.

III. Model Learning

For the estimation of the model's parameters, we consider the EM algorithm where we can calculate the posterior probability as following in the E-step:

$$\hat{Z}_{ij} = \frac{p_j \prod_{d=1}^{D} \phi_{ijd}}{\sum_{j=1}^{K} p_j \prod_{d=1}^{D} \phi_{ijd}} \tag{9}$$

where

$$\phi_{ijd} = \omega_d p\left(X_{id} \mid \xi_{jd}\right) + (1 - \omega_d) p\left(X_{id} \mid \lambda_d\right) \tag{10}$$

The parameters are estimated from the maximization of log-likelihood function, which can be written as:

$$\mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta) = \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij} \log\left(p\left(\vec{X}_i \mid \Theta_K\right)\right)$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\left\{\log p_j + \log\left[\omega_d p\left(\vec{X}_i \mid \xi_j\right) + (1-\omega_d)\, p\left(\vec{X}_i \mid \lambda\right)\right]\right\} \tag{11}$$

In the maximization step, the parameters can be estimated by taking the gradient of the log-likelihood in the previous equation with respect to each parameters, which gives the following for the mixing weights and the mean:

$$p_j^{new} = \frac{\sum_{i=1}^{N} h\left(j \mid \vec{X}_i, \Theta_M\right)}{N} \tag{12}$$

$$\mu_{jd}^{new} = \frac{\sum_{i=1}^{N} \frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right)\left\{\mathbf{X}_{id} - \frac{\int_{\partial_j} f(\mathbf{u}|\xi_j)\left(\mathbf{u}-\mu_{jd}\right)d\mathbf{u}}{\int_{\partial_j} f(\mathbf{u}|\xi_j)d\mathbf{u}}\right\}}{\sum_{i=1}^{N} \frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right)} \tag{13}$$

Note that in Eq. (13), the term $\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})d\mathrm{u}$ is the expectation of function $(\mathrm{u}-\mu_{jd})$ under the probability distribution $f(X_d|\xi_j)$. Then, this expectation can be approximated as:

$$\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}(s_{m_{jd}} - \mu_{jd})\mathrm{H}(s_{m_{jd}}|\Omega_j) \tag{14}$$

where $s_{m_{jd}} \sim f(\mathrm{u}|\xi_j)$ is a set of random variables drawn from the asymmetric Gaussian distribution for the particular component j of the mixture model. The term $\int_{\partial_j} f(\mathrm{u}|\xi_j)d\mathrm{u}$ in Eq. (13) can be approximated as:

$$\int_{\partial_j} f(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(s_{m_{jd}}|\Omega_j) \tag{15}$$

Thus, μ_{jd}^{new} can be written as:

$$\mu_{jd}^{new} = \frac{\sum_{i=1}^{N} \frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right)\left\{\mathrm{X}_{id} - \frac{\sum_{m=1}^{M}(s_{m_{jd}}-\mu_{jd})\mathrm{H}(s_{m_{jd}}|\Omega_j)}{\sum_{m=1}^{M}\mathrm{H}(s_{m_{jd}}|\Omega_j)}\right\}}{\sum_{i=1}^{N} \frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right)} \tag{16}$$

The left standard deviation can be estimated by maximizing the log-likelihood function with respect to $\sigma_{l_{jd}}$, which can be performed using Newton-Raphson method :

$$\sigma_{l_{jd}}^{new} = \sigma_{l_{jd}}^{old} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}^2}\right)^{-1}\left(\frac{\partial \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}}\right)\right] \tag{17}$$

where the first derivative of the model's complete data log-likelihood with respect to left standard deviation is given as follows:

$$\frac{\partial \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}} = \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id} \mid \xi_{jd}\right)}{\phi_{ijd}} \times h\left(j \mid \vec{X}_i, \theta_M\right)\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{l_{jd}}^3}\right) - \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id} \mid \xi_{jd}\right)}{\phi_{ijd} \times \sigma_{l_{jd}}^3} h\left(j \mid \vec{X}_i, \theta_M\right) \times \left\{\frac{\int_{\partial_j} \mathrm{g}_1(\mathrm{u} \mid \xi_j)(\mathrm{u}-\mu_{jd})^2 d\mathrm{u}}{\int_{\partial_j} \mathrm{g}_1(\mathrm{u} \mid \xi_j) d\mathrm{u}}\right\} \tag{18}$$

The term $\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 d\mathrm{u}$ can be approximated as below:

$$\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 du \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}} - \mu_{jd})^2 \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j) \tag{19}$$

where $\mathrm{l}_{m_{jd}} \sim \mathrm{g}_1(X_d|\xi_j)$ is a set of random variables drawn from the asymmetric Gaussian distribution with $\mathrm{u} < \mu_{jd}$ for the particular component j of the mixture model. Similarly, the term $\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)d\mathrm{u}$ in Eq. (18) can be approximated as:

$$\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j) \tag{20}$$

The same approximation for the second order derivative of the model's complete data log-likelihood with respect to left standard deviation is defined as follows:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}^2} &= -3\sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \gamma_{ij}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{l_{jd}}^4}\right) \\ &- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \gamma_{ij}\left(\frac{-2}{\sigma_{l_{jd}}^3\left(\sigma_{l_{jd}}+\sigma_{r_{jd}}\right)}\right) \times \left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{m_{jd}}-\mu_{jd}\right)^2 \mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \frac{\gamma_{ij}}{\sigma_{l_{jd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{m_{jd}}-\mu_{jd}\right)^4 \mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \frac{-3\gamma_{ij}}{\sigma_{l_{jd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{m_{jd}}-\mu_{jd}\right)^2 \mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N} \frac{\gamma_{ij}}{\sigma_{l_{jd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{m_{jd}}-\mu_{jd}\right)^2 \mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{m_{jd}} \mid \Omega_j\right)\right)^2}\right\}\end{aligned} \tag{21}$$

where

$$\gamma_{ij} = \frac{\omega_d p\left(X_{id} \mid \xi_{jd}\right)}{\phi_{ijd}} Z_{ij} \tag{22}$$

Similar approximations are used for the right standard deviation $\sigma_{r_{jd}}^{new}$:

$$\sigma_{r_{jd}}^{new} = \sigma_{r_{jd}}^{old} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{r_{jd}}^2}\right)^{-1}\left(\frac{\partial \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{r_{jd}}}\right)\right] \tag{23}$$

where

$$\frac{\partial \mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial \sigma_{r_{jd}}} = \sum_{i=1,\mathrm{X}_{id}\geq\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}}\times h\left(j\mid\vec{X}_i,\theta_M\right)\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma^3_{r_{jd}}}\right) - \sum_{i=1,\mathrm{X}_{id}\geq\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}\times\sigma^3_{l_{jd}}} h\left(j\mid\vec{X}_i,\theta_M\right)\times\left\{\frac{\int_{\partial_j}\ \mathrm{g}_2\left(\mathrm{u}\mid\xi_j\right)(\mathrm{u}-\mu_{jd})^2\,d\mathrm{u}}{\int_{\partial_j}\ \mathrm{g}_2\left(\mathrm{u}\mid\xi_j\right)d\mathrm{u}}\right\} \tag{24}$$

The term $\int_{\partial_j} \mathrm{g}_2(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 d\mathrm{u}$ can be approximated as below:

$$\int_{\partial_j} \mathrm{g}_2(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j) \tag{25}$$

where $\mathrm{r}_{m_{jd}} \sim \mathrm{g}_2(X_d|\xi_j)$ is a set of random variables drawn from the asymmetric Gaussian distribution with $\mathrm{u} \geq \mu_{jd}$ for the particular component j of the mixture model. Similarly, the term $\int_{\partial_j} \mathrm{g}_2(\mathrm{u}|\xi_j)d\mathrm{u}$ in Eq. (24) can be approximated as:

$$\int_{\partial_j} \mathrm{g}_2(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j) \tag{26}$$

Similar approximations are used for $\frac{\partial^2\mathcal{L}(\mathcal{X},\mathcal{Z}|\Theta)}{\partial\sigma^2_{r_{jd}}}$ is given as following:

$$\begin{aligned}\frac{\partial^2 \mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma^2_{r_{jd}}} &= -3\sum_{X_{id}\geq\mu_{jd}}^{N}\gamma_{ij}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma^4_{r_{jd}}}\right) - \sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\gamma_{ij}\left(\frac{-2}{\sigma^3_{r_{jd}}\left(\sigma_{r_{jd}}+\sigma_{r_{jd}}\right)}\right)\times\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma^6_{r_{jd}}}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^4\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\frac{-3\gamma_{ij}}{\sigma^4_{r_{jd}}}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\} \\ &- \sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma^6_{r_{jd}}}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)\right)^2}\right\}\end{aligned} \tag{27}$$

The parameters of background Gaussian can be estimated using the following equations:

$$\eta_d^{new} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j\mid\vec{X}_i,\theta_M\right)\right]X_{id}}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j\mid\vec{X}_i,\theta_M\right)} \tag{28}$$

$$\delta_d^{2^{\text{new}}} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j\mid\vec{X}_i,\theta_M\right)\right](X_{id}-\eta_d)^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j\mid\vec{X}_i,\theta_M\right)} \tag{29}$$

$$\omega_d^{\text{new}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p(X_{id}|\xi_{jd})}{\phi_{ijd}}h\left(j\mid\vec{X}_i,\theta_M\right)}{N} \tag{30}$$

A. Model selection via MML and complete algorithm

In order to estimate the number of components of the mixture model, we apply MML criterion which consists of minimizing the message length given by the following equation

$$\begin{aligned}\text{MessLens} \approx &-\log p\left(\Theta_M\right)+\frac{c}{2}\left(1+\log\frac{1}{12}\right) \\ &+\frac{1}{2}\log\left|I\left(\Theta_M\right)\right|-\log p\left(\mathcal{X}\mid\Theta_M\right)\end{aligned} \tag{31}$$

where $p\left(\Theta_M\right)$ is prior distribution, $I\left(\Theta_M\right)$ denotes the Fisher information matrix, $\log p\left(\mathcal{X}\mid\Theta_M\right)$ is log-likelihood. Here the constant value c represents the total number of parameters, which is equal $M+D+3DM+2D$, $\left|I\left(\Theta_M\right)\right|$ denotes the determinant of the Fisher information matrix of our model which is very hard to calculate analytically, so we assume that each group of parameters is independent, which allows the factorization of $p\left(\Theta_M\right)$ and $I\left(\Theta_M\right)$. Moreover, we adopt the uninformative Jeffrey's prior for each group of parameters as prior distributions without knowing the parameters. Then, we have the following equation:

$$\begin{aligned}\text{MessLens} \approx &\frac{c}{2}\left(1+\log\frac{1}{12}\right)+\frac{c}{2}(\log N)+\frac{3M}{2}\sum_{d=1}^{D}\log\omega_d \\ &+\frac{3D}{2}\sum_{j=1}^{M}\log p_j+\sum_{d=1}^{D}\log\left(1-\omega_d\right)-\log p\left(\mathcal{X}\mid\theta_M\right)\end{aligned} \tag{32}$$

The minimization of the previous equation gives the following:

$$p_j^* = \frac{\max\left(\sum_{i=1}^{N}h\left(j\mid\vec{X}_i,\Theta_M\right)-\frac{3D}{2},0\right)}{\sum_{j=1}^{M}\max\left(\sum_{i=1}^{N}h\left(j\mid\vec{X}_i,\Theta_M\right)-\frac{3D}{2},0\right)} \tag{33}$$

$$\omega_d^* = \frac{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M}a_{ijd}-\frac{3M}{2},0\right)}{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M}U_{ijd}-\frac{3M}{2},0\right)+\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M}V_{ijd}-1,0\right)} \tag{34}$$

where

$$U_{ijd} = h\left(j\mid\vec{X}_i,\Theta_M\right)\frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}} \tag{35}$$

$$V_{ijd} = h\left(j\mid\vec{X}_i,\Theta_M\right)\frac{\left(1-\omega_d\right)p\left(X_{id}\mid\lambda_d\right)}{\phi_{ijd}} \tag{36}$$

The complete learning of BAGGM-FS is given in Algorithm 1, where t_{min} is the minimum threshold used to monitor the log-likelihood convergence, $epoch_{max}$ is maximum number of iterations, K_{min} and K_{max} define the searching range for the optimal number of clusters. In the initialization step, K-Means is used to initialize the parameters of each clusters.

Algorithm 1 Feature Selection for BAGMM

Input:Dataset $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$, t_{min}, $epoch_{max}$, K_{min}, K_{max}.
Output: Θ, $\mathcal{Z}$, K^*.
for $K_{min} \leq K \leq K_{max}$ do
  {Initialization}:
  K-Means algorithm (Compute $\vec{\mu}_1, \ldots, \vec{\mu}_K$ & cluster assignment)
  Set $\vec{\omega}$ = 0.5
  for all $1 \leq j \leq K$ do
    Computation of p_j and $\{\vec{\mu}_j = \vec{\mu}_j, (\vec{\sigma}_{l_j}$ & $\vec{\sigma}_{r_j})$=$\vec{\sigma}_j\}$ and $\vec{\lambda} = \{\vec{\eta} = \vec{\mu}_j, \vec{\delta} = \vec{\sigma}_j\}$
  {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 $h\left(j \mid \vec{X}_i, \Theta_M\right)$ for $i = 1, \ldots, N$ using Eq. (9).
    {[**M step**]}:
    update bounded support range
    for all $1 \leq j \leq K$ do
      Estimate p_j, $\vec{\mu}_j$, $\vec{\sigma}_{l_j}$ & $\vec{\sigma}_{r_j}$ using Eqs. (12, 16, 17, & 23).
    end for
    Estimate $\vec{\eta}$, $\vec{\delta}$ & $\vec{\omega}$ using Eqs. (28, 29, & 30).
    If $p_j = 0$, jth cluster is pruned
    If $\omega_d = 0$, $p\left(X_{id} \mid \xi_{jd}\right)$ is pruned
    If $\omega_d = 1$, $p\left(X_{id} \mid \lambda_d\right)$ is pruned
  end while
  Compute $K^* = \arg\min MML(K)$ using Eq. (32)
end for

IV. Experimental Results

In this section, the effectiveness of our model is tested on several real-world applications, including human gender recognition, human activity categorization, and human part recognition. We have compared our approach (BAGMM-FS) with bounded asymmetric Gaussian mixture model (BAGMM), asymmetric Gaussian mixture model (AGMM), asymmetric Gaussian mixture model with feature selection (AGMM-FS), Gaussian mixture model (GMM), and bounded generalized Gaussian mixture model (BGGMM). For comparison, we use the following clustering metrics: accuracy, which is computed as: $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$, precision, which is computed as: $\left(\frac{TP}{TP+FP}\right)$, recall, which is computed as: $\left(\frac{TP}{TP+FN}\right)$, F1 Score, which is computed as: $2*(precision*recall)/(precision+recall)$. Here, the term TP stands for true positives, TN for true negatives, FP for false positives, and FN stands for false negatives. In addition, we use the silhouette score [23] which indicates the overlapping clusters with the range from -1 to 1, and 1 is the best value, -1 for the worst value, value near 0 indicates overlapping clusters. It is only defined if the number of clusters is greater than 2. So if all the data instances are assigned to one cluster, the silhouette score is not applicable, and it will be denoted by N/A. Finally, we consider the classification entropy (CE) index [24], which indicates good clustering when it is low and poor clustering when it is high.

A. Human Activity Categorization

Human activity categorization (HAR) has received a lot of research attention in the last decade [25], [26]. It has numerous practical applications such as surveillance and health monitoring. In this section, we consider a human activity categorization dataset called UCI Daily and Sports Activity dataset (DSAD)[1] for our experiment [27]. It contains 19 different kinds of signal data, acquired from different sensors, of activities recorded in a flat outdoor area on campus, such as sitting, standing, etc., performed by eight subjects (4 female, 4 male, between the ages 20 and 30). In our simulations, firstly, eight daily activities from the first subject, including sitting, standing, walking, jumping, playing basketball, rowing, exercising, and running, are chosen to be classified to prove our mixture model's effectiveness. There are 992 observations with 45 dimensions in 8 clusters. As we can see from the results in the Table I, the mixture models with feature selection outperform the other models, which demonstrates the effectiveness of feature selection for high-dimensional data. GMM, the baseline of mixture models, has the lowest accuracy. Note that our proposed model outperforms all other models with respect to all the calculated metrics and we have received very high accuracy of 96.47% for this experiment. In addition, our model has converged with fewer epochs than AGMM and AGMM-FS under the same initialization method and learning rate. We have also compared the MML for BAGMM-FS with the MML for BAGMM, AGMM-FS, and AGMM in Fig. 1. According to this figure only BAGMM and BAGMM-FS were able to find the correct number of components which is 8, while AGMM and AGMM-FS favored 10 clusters.

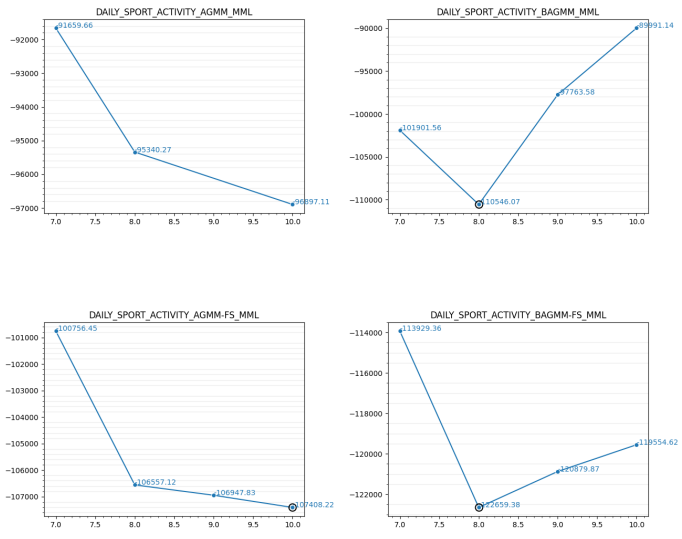

Fig. 1: MML for the activities clustering application using different mixture models.

[1]**DSAD** dataset available at: http://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities

TABLE I: 8 common daily activities, from the first subject, clustering using different mixture models.

Models	Time	Epoch	Accuracy	Precision	Recall	F1-score	Silhouette	CE
BAGMM-FS	3.11	3	96.47%	97.25%	96.47%	96.40%	0.451	0.004
BAGMM	3.04	1	95.56%	96.33%	95.56%	95.29%	0.454	1.49
AGMM-FS	2.67	38	96.37%	97.18%	96.37%	96.29%	0.451	0.002
AGMM	0.468	12	95.86%	96.89%	95.86%	95.75%	0.452	0.002
BGGMM	494.95	7	95.56%	96.33%	95.56%	95.30%	0.454	4.17e-7
GGMM	0.640	7	62.5%	43.81%	62.50%	50.03%	0.198	0.430
GMM	0.014	1	44.76%	36.74%	44.75%	34.83%	0.211	0.641

For another experiment with this dataset, we cluster different sitting activities from the 8 subjects representing by 992 data instances in total with 45 dimensions. From Table II, we can see again that feature selection improves the clustering results. Mixture models without feature selection have almost the same accuracy as the baseline GMM, which is around 71%. Note that our proposed mixture model distinguishes itself as compared to the other mixture models with respect to all the considered clustering metrics.

B. Gender Recognition

Gender recognition is an important task in computer vision and has received increasing attention with the rapid development of machine learning. There are numerous applications that require gender recognition like human-computer interaction, image-based indexing and searching, biometrics, and even targeted advertising. Some studies show that a human can classify between a male and a female simply (over 95% accuracy from faces [28]). However, it's a complex task for machines because of people's variation status at different light intensities, such as different postures, angles, etc [29]. Without prior information about training data, mixture models as the unsupervised learning method can be effective for gender recognition. In this section, we will verify BAGMM-FS on three well-known datasets, PARSE-27k dataset [2], PETA dataset [3] and Human attribute dataset[4] [30]–[33]. Fig. 2 shows sample images for gender recognition in PARSE-27k dataset. Compared with other human attribute datasets, the PARSE-27k dataset has relatively more minor variance because it only contains crops of pedestrian bounding boxes obtained by a pedestrian detector. For simplicity, the website of PARSE-27k provides HDF5 file format of 64 $\times$ 128 sized crops, including labels for quick experiments, so we did not need to crop images by ourselves. The PETA dataset consists of 19,000 images annotated with 61 binary and 4 multi-class attributes. The PETA dataset comprises 10 sub-datasets, including CUHK, CAVIAR4REID, and MIT, recorded at different places with different camera angles and viewpoints. In this experiment, we choose the CUHK sub-dataset with resolutions of 80 $\times$ 160, a high camera angle, and a varying viewpoint. Fig. 3 shows sample images from the PETA dataset.

[2]**PARSE-27k** dataset available at: https://www.vision.rwth-aachen.de/page/parse27k

[3]**PETA** dataset available at: http://mmlab.ie.cuhk.edu.hk/projects/PETA.html

[4]**Human attribute** dataset available at: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/poselets/

Fig. 2: Samples images from PARSE-27k dataset.

Fig. 3: Samples images from CUHK sub-dataset in PETA dataset

In order to describe the images, we have considered bag of visual words (BOVW) [34]. The basic idea is to extract local features for each image using scale invariant feature transform (SIFT) [35]. Then, K-Means is used to cluster the 128-dimensional descriptors for building the visual words vocabulary, where size is equal to the number of centroids. In short, the BOVW works by extracting features such as shape, texture, etc., in a dense grid of rectangular windows and constructs a fixed-size visual vocabulary by counting each visual word's occurrence in an image.

TABLE II: Clustering of the sitting activities of the 8 subjects using different mixture models.

Mixture Models	Time	Epoch	Accuracy	Precision	Recall	F1-score	Silhouette	CE
BAGMM-FS	5.87	6	90.52%	93.13%	90.52%	89.80%	0.514	0.009
BAGMM	2.23	2	72.78%	70.88%	72.78%	71.47%	0.569	0.011
AGMM-FS	0.641	9	84.97%	78.35%	84.97%	80.43%	0.593	0.156
AGMM	0.105	1	72.47%	63.37%	72.47%	65.58%	0.624	0.384
BGGMM	107.74	16	72.47%	71.41%	72.48%	71.50%	0.495	4.4e-77
GGMM	0.237	3	72.47%	63.37%	72.47%	65.58%	0.624	0.384
GMM	0.015	1	71.67%	59.87%	71.67%	63.67%	0.556	0.383

Regarding the PARSE-27k experiment, we selected 2,000 images, composed of 1,000 female photos and 1,000 male photos, by considering a visual vocabulary having a size of 110. Besides, the distribution of clusters is so imbalanced which makes the clustering task very challenging. The clustering results using different mixtures are summarized in Table III and show clearly that our model outperformed all the others. For another gender recognition experiment, we choose 346 images (200 for males and 146 females) from the CUHK folder in the PETA dataset. We also employ SIFT and BOVW approach to extract feature vectors from these images. We considered a vocabulary with a size of 130 after many tries. The clustering results for this data set are given in Table III and we can see again that our model has an excellent performance as compared to the other models.

We have also considered a challenging dataset called Human attribute dataset [32], [33]. The Human attribute dataset of H3D folder comprises 750 images in total (437 for male images, 313 for female images) in which there are nine attributes and visible bounding boxes of person for each image. The attribute value is 1 if it is present, -1 if it is not, and 0 if it is unspecified which we have not considered nor use in our experiments. The same feature extraction process used above is also considered for this data set.

Fig. 4: Samples images from human attribute dataset.

This dataset is different from the datasets mentioned above because of its complex and colorful backgrounds. We randomly picked up 313 images (half males and half females). After feature extraction, we considered a vocabulary of size 100. It's observed that our proposed model performs better than the other mixture models, as shown in Table III. In particular, we can observe that several silhouette score values are N/A, which means that the associated models failed to distinguish both classes. Our proposed BAGMM-FS has the highest accuracy of 70.61% as compared with BAGMM with 62.77% accuracy and fewer iterations as compared with AGMM-FS because of bounded support. Note that feature selection can help mixture models converge faster observed from the execution time of AGMM and AGMM-FS.

V. Conclusion

We propose a statistical framework for simultaneous clustering and feature selection based on BAGMM. The proposed statistical model is learned using the EM algorithm to estimate the mixture's parameters and select the number of clusters by MML. In contrast with other dimension reduction approaches, our proposed algorithm uses the full dimensionality of the data and gives a weight to each feature automatically. Using two applications that involve human activity and gender recognition, we have shown that the proposed model outperforms other mixture models considered for comparison. It is demonstrated through several performance measures. For two experiments on human activity recognition, our model has achieved a very high accuracy of 96.47% and 90.52%, demonstrating the proposed algorithm's effectiveness. Future works could be devoted to applying the proposed framework to other challenging applications or considering other learning techniques such as Bayesian inference or variational Bayes.

References

[1] G. J. McLachlan, S. X. Lee, and S. I. Rathnayake, "Finite mixture models," *Annual Review of Statistics and Its Application*, vol. 6, no. 1, pp. 355–378, 2019. [Online]. Available: https://doi.org/10.1146/annurev-statistics-031017-100325

[2] N. Bouguila and W. Fan, *Mixture Models and Applications*. Springer, 2020.

[3] 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. 1 – 13, 2008. [Online]. Available: https://doi.org/10.1117/1.2898125

[4] T. Elguebaly and N. Bouguila, "Bayesian learning of finite generalized gaussian mixture models on images," *Signal Processing*, vol. 91, no. 4, pp. 801–820, 2011.

[5] 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.

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

[7] M. Azam and N. Bouguila, "Bounded Generalized Gaussian Mixture Model with ICA," *Neural Processing Letters*, vol. 49, no. 3, pp. 1299–1320, Jun 2019. [Online]. Available: https://doi.org/10.1007/s11063-018-9868-7

TABLE III: Gender recognition results.

Models	dataset	Time	Epoch	Accuracy	Precision	Recall	F1-score	Silhouette	CE
BAGMM-FS	PETA	2.18	7	**81.21%**	81.52%	81.21%	81.29%	0.018	0.170
BAGMM	PETA	1.322	4	51.44%	48.73%	51.44%	48.97%	-0.002	0.011
AGMM-FS	PETA	0.813	45	57.80%	33.41%	57.80%	42.34%	N/A	0.693
AGMM	PETA	0.237	21	57.80%	33.41%	57.80%	42.34%	N/A	0.693
BGGMM	PETA	15.93	3	39.59%	41.91%	39.59%	36.31%	0.075	0.011
GGMM	PETA	2.046	300	57.80%	33.41%	57.80%	42.34%	N/A	0.693
GMM	PETA	0.024	1	57.80%	33.41%	57.80%	42.34%	N/A	0.693
BAGMM-FS	PARSE-27k	3.13	5	**77.33%**	82.49%	77.33%	67.83%	-0.122	0.005
BAGMM	PARSE-27k	2.02	4	50.18%	82.47%	50.18%	51.47%	0.055	0.039
AGMM-FS	PARSE-27k	4.10	13	76.93%	59.19%	76.93%	66.90%	N/A	0.693
AGMM	PARSE-27k	37.61	209	76.93%	59.19%	76.93%	66.90%	N/A	0.693
BGGMM	PARSE-27k	30.33	6	70.61%	76.80%	70.61%	72.52%	0.012	0.117
GGMM	PARSE-27k	1.101	3	76.93%	59.19%	76.93%	66.90%	N/A	0.693
GMM	PARSE-27k	0.112	1	76.93%	59.19%	76.93%	66.90%	N/A	0.693
BAGMM-FS	Human attribute	0.506	1	**70.61%**	73.89%	70.61%	69.56%	0.125	0.116
BAGMM	Human attribute	0.504	1	62.77%	78.66%	62.77%	56.79%	0.110	0.007
AGMM-FS	Human attribute	0.571	16	50.00%	25.00%	50%	33.33%	N/A	0.693
AGMM	Human attribute	4.003	300	50.00%	25.00%	50%	33.33%	N/A	0.693
BGGMM	Human attribute	48.073	8	61.98%	62.43%	61.98%	61.62%	0.097	0.009
GGMM	Human attribute	0.121	3	50.00%	25.00%	50%	33.33%	N/A	0.693
GMM	Human attribute	0.038	1	50.00%	25.00%	50%	33.33%	N/A	0.693

[8] A. Hyvärinen and P. Hoyer, "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[9] 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.

[10] N. Nacereddine, S. Tabbone, D. Ziou, and L. Hamami, "Asymmetric generalized gaussian mixture models and em algorithm for image segmentation," in *2010 20th International Conference on Pattern Recognition*, 2010, pp. 4557–4560.

[11] T. M. Nguyen, Q. J. Wu, D. Mukherjee, and H. Zhang, "Bounded asymmetric mixture model for medical image segmentation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 1031–1035.

[12] 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

[13] G. Mclachlan and T. Krishnan, *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*, 01 2007.

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

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

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

[17] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, Sept 2004.

[18] A. Jain and D. Zongker, "Feature selection: evaluation, application, and small sample performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 2, pp. 153–158, 1997.

[19] N. Bouguila, K. Almakadmeh, and S. Boutemedjet, "A finite mixture model for simultaneous high-dimensional clustering, localized feature selection and outlier rejection," *Expert Systems with Applications*, vol. 39, no. 7, pp. 6641–6656, 2012.

[20] N. Bouguila, D. Ziou, and S. Boutemedjet, "Simultaneous Non-gaussian Data Clustering, Feature Selection and Outliers Rejection," in *Pattern Recognition and Machine Intelligence*, S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, and S. K. Pal, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 364–369.

[21] A. Cord, C. Ambroise, and J.-P. Cocquerez, "Feature selection in robust clustering based on laplace mixture," *Pattern Recognition Letters*, vol. 27, no. 6, pp. 627–635, 2006.

[22] T. Elguebaly and N. Bouguila, "Simultaneous high-dimensional clustering and feature selection using asymmetric Gaussian mixture models ," *Image and Vision Computing*, vol. 34, pp. 27 – 41, 2015.

[23] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[24] F. Iglesias, T. Zseby, and A. Zimek, "Absolute cluster validity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2096–2112, 2020.

[25] Z. Chen, L. Zhang, Z. Cao, and J. Guo, "Distilling the knowledge from handcrafted features for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4334–4342, 2018.

[26] A. Ignatov, "Real-time human activity recognition from accelerometer data using convolutional neural networks," *Applied Soft Computing*, vol. 62, pp. 915–922, 2018.

[27] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," *Pattern Recognition*, vol. 43, no. 10, pp. 3605–3620, 2010.

[28] V. Bruce, A. M. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney, "Sex discrimination: how do we tell the difference between male and female faces?" *perception*, vol. 22, no. 2, pp. 131–152, 1993.

[29] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3633–3645, 2014.

[30] P. Sudowe, H. Spitzer, and B. Leibe, "Person Attribute Recognition with a Jointly-trained Holistic CNN Model," in *ICCV'15 ChaLearn Looking at People Workshop*, 2015.

[31] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 789–792.

[32] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *International Conference on Computer Vision*, sep 2009. [Online]. Available: http://www.eecs.berkeley.edu/~lbourdev/poselets

[33] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1543–1550.

[34] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *BMVC 2010-21st British Machine Vision Conference*, 2010.

[35] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.