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

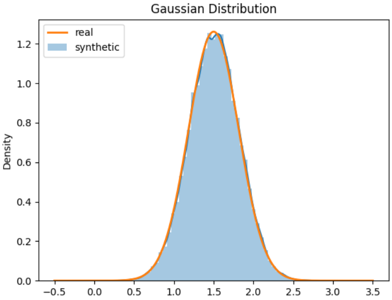
IEEE IRI February 23, 2022

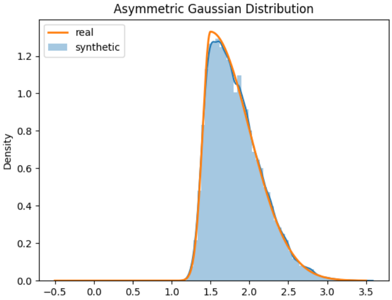
Outline

1. **Proposed Model**
   - Feature Selection
2. **Model Learning**
   - E Step
   - M Step
   - Model selection via MML
3. **Experimental Results**
   - Clustering Metrics
   - Human Activity Categorization
   - Gender Recognition
4. **Conclusion**

Université Concordia University

Asymmetric Gaussian Distribution



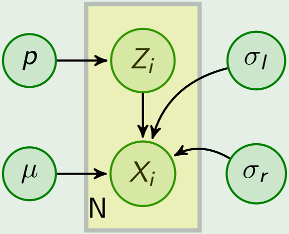


- ☞ Gaussian distribution (GD) assumes that the data is symmetric and has an infinite range, which prevents it from having a good modeling capability in the presence of outliers.
- ☞ Asymmetric Gaussian distribution (AGD) has been proposed to model asymmetric real-world data by having two variance parameters controlling the left and right parts of the distribution.

Proposed Model

Mixture of Asymmetric Gaussian Distributions (AGMM)

AGMM



Graphical representation of an asymmetric Gaussian mixture model

Mathmatical Definition

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

- $p\left(\vec{X}_i|\xi_j\right)$ is the PDF of AGD
- Z_{ij} is the hidden variable which satisfy $Z_{ij} \in \{0, 1\}$
- p_j are the mixing weights that satisfy $p_j \geq 0$, $\sum_{j=1}^{K} p_j = 1$.

Mixture of Bounded Asymmetric Gaussian Distributions

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

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

where $\xi_j = (\vec{\mu}_j, \vec{\sigma_{l_j}}, \vec{\sigma_{r_j}})$ 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. $f(\vec{X}|\xi_j)$ is the PDF of AGD.

the term $\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j)d\mathrm{u}$ in Eq. (2) is the normalized constant that shows the share of $f(\vec{X}|\xi_j)$ which belongs to the support region ∂.

Feature Selection

We can get the complete data log-likelihood by taking the logarithm of Eq. (1) as follows.

$$\log p(\mathcal{X}, Z \mid \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \log \left[p\left(\vec{X}_i \mid \xi_j\right) p_j \right] \tag{4}$$

According to Eq. (4), 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.

Feature Selection

$$p(\vec{X}_i \mid \Theta, \vec{\lambda}, \vec{\varphi}) = \sum_{j=1}^{K} p_j \prod_{d=1}^{D} p\left(X_d \mid \xi_{jd}\right)^{\varphi_d} p\left(X_d \mid \lambda_d\right)^{1-\varphi_d} \quad (5)$$

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

According to [1], we can obtain:

$$p\left(\vec{X}_i \mid \Theta_K\right) = \sum_{j=1}^{K} p_j \prod_{d=1}^{D} \left[\omega_d p\left(X_d \mid \xi_{jd}\right) + \left(1 - \omega_d\right) p\left(X_d \mid \lambda_d\right)\right] \quad (6)$$

Where $p\left(X_d \mid \lambda_d\right)$ is the background Gaussian distribution with parameters $\vec{\lambda} = \{\vec{\eta}, \vec{\delta}\}$. 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 $0 \le \omega_d \le 1$, $d = 1, \ldots, D$.

E Step

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:

$$\hat{Z}_{ij} = \frac{p_j \prod_{d=1}^{D} \phi_{ijd}}{\sum_{j=1}^{K} p_j \prod_{d=1}^{D} \phi_{ijd}} \quad (7)$$

where

$$\phi_{ijd} = \omega_d \, p\left(X_{id} \mid \xi_{jd}\right) + \left(1 - \omega_d\right) p\left(X_{id} \mid \lambda_d\right) \quad (8)$$

Mixing Parameter Estimation

The log-likelihood function can be written as:

$$\mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \log \left(p \left(\vec{X}_i \mid \Theta_K \right) \right)$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \left\{ \log p_j + \log \left[\omega_d p \left(\vec{X}_i \mid \xi_j \right) + (1 - \omega_d) p \left(\vec{X}_i \mid \lambda \right) \right] \right\} \quad (9)$$

In M-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:

Estimation of p

$$p_j^{new} = \frac{\sum_{i=1}^{N} h \left(j \mid \vec{X}_i, \Theta_M \right)}{N} \quad (10)$$

Mean Parameter Estimation

Estimation of μ

$$\mu_{jd}^{new} = \frac{\sum_{i=1}^{N} \frac{\omega_d p(X_{id}|\xi_{jd})}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right) \left\{ \mathrm{X}_{id} - \frac{\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})du}{\int_{\partial_j} f(\mathrm{u}|\xi_j)du} \right\}}{\sum_{i=1}^{N} \frac{\omega_d p(X_{id}|\xi_{jd})}{\phi_{ijd}} h\left(j \mid \vec{X}_i, \Theta_M\right)} \quad (11)$$

In Eq. (11), the term $\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u} - \mu_{jd})du$ 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:

$$\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u} - \mu_{jd})du \approx \frac{1}{M} \sum_{m=1}^{M} (s_{m_{jd}} - \mu_{jd}) \mathrm{H}(s_{m_{jd}}|\Omega_j) \quad (12)$$

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.

Left And Right Standard Deviation Estimation

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

Estimation of σ_l and σ_r

$$\sigma_{l_{jd}}^{new} = \sigma_{l_{jd}}^{old} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{l_{jd}}} \right) \right] \tag{13}$$

$$\sigma_{r_{jd}}^{new} = \sigma_{r_{jd}}^{old} - \left[\left(\frac{\partial^2 \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{r_{jd}}^2} \right)^{-1} \left(\frac{\partial \mathcal{L}(\mathcal{X}, \mathcal{Z} \mid \Theta)}{\partial \sigma_{r_{jd}}} \right) \right] \tag{14}$$

The details are covered in the paper.

Parameters of Background Gaussian Estimation

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

$$\eta_d^{new} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i,\theta_M\right)\right]X_{id}}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i,\theta_M\right)} \quad (15)$$

$$\delta_d^{2new} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i,\theta_M\right)\right](X_{id}-\eta_d)^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i,\theta_M\right)} \quad (16)$$

$$\omega_d^{new} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p(X_{id}|\xi_{jd})}{\phi_{ijd}}h\left(j \mid \vec{X}_i,\theta_M\right)}{N} \quad (17)$$

Model selection criteria

EM algorithm requires an appropriate number of clusters found by model selection criteria.

- stochastic (e.g. Markov Chain Monte Carlo)
- deterministic approaches including Akaike's information criterion (AIC) [2], Schwarz's Bayesian information criterion (BIC) [3], the Laplace empirical criterion (LEC) [4] and minimum message length (MML) [1, 5], et
- re-sampling methods

The MML has been shown to have better performance among most model selection criteria in most cases.

Minimum Message Length

MML

$$\text{MessLens} \approx -\log p(\Theta_M) + \frac{c}{2}\left(1 + \log\frac{1}{12}\right) + \frac{1}{2}\log|I(\Theta_M)| - \log p(\mathcal{X} \mid \Theta_M) \quad (18)$$

- $p(\Theta_M)$ is prior distribution.
- $I(\Theta_M)$ denotes the Fisher information matrix.
- $\log p(\mathcal{X} \mid \Theta_M)$ is log-likelihood.
- c represents the total number of parameters, which is equal $M + D + 3DM + 2D$
- $|I(\Theta_M)|$ denotes the determinant of the Fisher information matrix.

Minimum Message Length

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.

MML can be rewritten as

$$
\begin{aligned}
\text{MessLens} \approx & \frac{c}{2}\left(1+\log\frac{1}{12}\right)+\frac{c}{2}(\log N)+\frac{3M}{2}\sum_{d=1}^{D}\log\omega_d \\
& +\frac{3D}{2}\sum_{j=1}^{M}\log p_j+\sum_{d=1}^{D}\log\left(1-\omega_d\right)-\log p\left(\mathcal{X}\mid\theta_M\right)
\end{aligned} \tag{19}
$$

Minimum Message Length

The minimization of the Eq. (19) gives the following:

$$p_j^* = \frac{\max\left(\sum_{i=1}^{N} h\left(j \mid \vec{X}_i, \Theta_M\right) - \frac{3D}{2}, 0\right)}{\sum_{j=1}^{M} \max\left(\sum_{i=1}^{N} h\left(j \mid \vec{X}_i, \Theta_M\right) - \frac{3D}{2}, 0\right)} \quad (20)$$

$$\omega_d^* = \frac{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} a_{ijd} - \frac{3M}{2}, 0\right)}{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} U_{ijd} - \frac{3M}{2}, 0\right) + \max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} V_{ijd} - 1, 0\right)} \quad (21)$$

where

$$U_{ijd} = h\left(j \mid \vec{X}_i, \Theta_M\right) \frac{\omega_d\, p\left(X_{id} \mid \xi_{jd}\right)}{\phi_{ijd}} \quad (22)$$

$$V_{ijd} = h\left(j \mid \vec{X}_i, \Theta_M\right) \frac{\left(1 - \omega_d\right) p\left(X_{id} \mid \lambda_d\right)}{\phi_{ijd}} \quad (23)$$

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 & $\vec{\omega} = 0.5$)
  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 $i = 1, \ldots, N$
    {[**M step**]}:
    update bounded support range
    for all $1 \leq j \leq K$ do
      Estimate p_j, $\vec{\mu}_j$, $\vec{\sigma}_{l_j}$ & $\vec{\sigma}_{r_j}$
    Estimate $\vec{\eta}$, $\vec{\delta}$ & $\vec{\omega}$
    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
  Compute $K^* = \arg\min MML(K)$

Clustering Metrics

- Accuracy is computed as: $\left(\frac{TP+TN}{TP+TN+FP+FN}\right)$
- Precision is computed as: $\left(\frac{TP}{TP+FP}\right)$
- Recall is computed as: $\left(\frac{TP}{TP+FN}\right)$
- F1 Score is computed as: $2 * (precision * recall)/(precision + recall)$
- Silhouette score 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.
- Classification entropy (CE) index indicates good clustering when it is low and poor clustering when it is high.

the term *TP* stands for true positives, *TN* for true negatives, *FP* for false positives, and *FN* stands for false negatives

Human Activity Categorization[1]

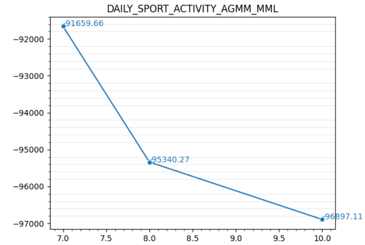
We consider a human activity categorization dataset called UCI Daily and Sports Activity dataset (DSAD)[1], which contains 19 different kinds of signal data performed by eight subjects (4 female, 4 male)

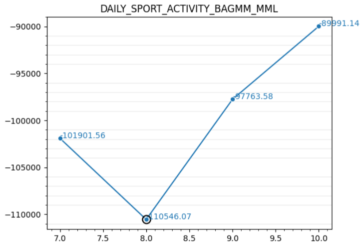
Eight daily activities from the first subject, including sitting, standing, walking, jumping, playing basketball, rowing, exercising, and running, are chosen to be classified.

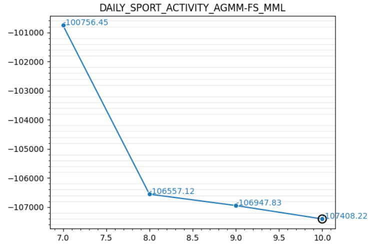
Table: 8 common daily activities, from the first subject, clustering using different mixture models.

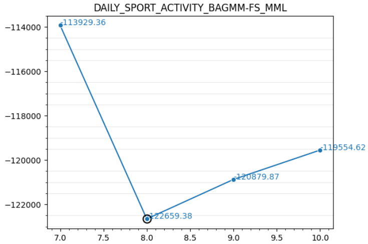
Models	Time	Epoch	Accuracy	Precision	Recall	F1-score	Silhouette	CE
BAGMM-FS	3.11	3	96.47%	97.25%	96.47%	96.40%	0.451	0.004
BAGMM	3.04	1	95.56%	96.33%	95.56%	95.29%	0.454	1.49
AGMM-FS	2.67	38	96.37%	97.18%	96.37%	96.29%	0.451	0.002
AGMM	0.468	12	95.86%	96.89%	95.86%	95.75%	0.452	0.002
BGGMM	494.95	7	95.56%	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

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









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.

Human Activity Categorization[2]

We also cluster different sitting activities from the 8 subjects in this dataset. **Feature selection improves the clustering results.** Mixture models without feature selection have almost the same accuracy as the baseline GMM. Our proposed model distinguishes itself as compared to the other mixture models with respect to all considered clustering metrics.

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

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

We verify BAGMM-FS on three well-known datasets, PARSE-27k dataset (P27K), PETA dataset and Human attribute dataset (H3D).

Using bag of visual words (BOVW) to describe the images:

1. extract local features for each image using scale invariant feature transform (SIFT)
2. apply K-Means to cluster the 128-dimensional descriptors for building the visual words vocabulary

Figure: Samples images from datasets.

Table: Gender recognition results.

Models	dataset	Time	Epoch	Acc	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	P27K	3.13	5	**77.33**	82.49%	77.33%	67.83%	-0.122	0.005
BAGMM	P27K	2.02	4	50.18	82.47%	50.18%	51.47%	0.055	0.039
AGMM-FS	P27K	4.10	13	76.93	59.19%	76.93%	66.90%	N/A	0.693
AGMM	P27K	37.61	209	76.93	59.19%	76.93%	66.90%	N/A	0.693
BGGMM	P27K	30.33	6	70.61	76.80%	70.61%	72.52%	0.012	0.117
GGMM	P27K	1.101	3	76.93	59.19%	76.93%	66.90%	N/A	0.693
GMM	P27K	0.112	1	76.93	59.19%	76.93%	66.90%	N/A	0.693
BAGMM-FS	H3D	0.506	1	**70.61**	73.89%	70.61%	69.56%	0.125	0.116
BAGMM	H3D	0.504	1	62.77	78.66%	62.77%	56.79%	0.110	0.007
AGMM-FS	H3D	0.571	16	50.00	25.00%	50%	33.33%	N/A	0.693
AGMM	H3D	4.003	300	50.00	25.00%	50%	33.33%	N/A	0.693
BGGMM	H3D	48.073	8	61.98	62.43%	61.98%	61.62%	0.097	0.009
GGMM	H3D	0.121	3	50.00	25.00%	50%	33.33%	N/A	0.693
GMM	H3D	0.038	1	50.00	25.00%	50%	33.33%	N/A	0.693

Conclusion

- Proposed a statistical framework for simultaneous clustering and feature selection based on BAGMM
- Using EM algorithm to estimate the mixture's parameters and select the number of clusters by MML
- Applying full dimensionalities of the data and giving a weight to each feature automatically
- Validated using two applications that involve human activity and gender recognition
- Compared with other mixture models
- Demonstrated effectiveness of our model through several performance measures

For Further Reading I

Zixiang Xian, Muhammad Azam, Nizar Bouguila
Statistical Modeling Using Bounded Asymmetric Gaussian Mixtures: Application to Human Action and Gender Recognition
IEEE 22nd International Conference on Information Reuse and Integration for Data Science, *IEEE IRI 2021*

References I

- 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, pp. 1154–1166, Sept 2004.
- H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, December 1974.
- G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.
- G. McLachlan and D. Peel, "Finite mixture models.,(john wiley & sons: New york.)," 2000.

References II

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.

Q&A