Generative Models Based on the Bounded Asymmetric Gaussian Distribution

Zixiang Xian

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Information Systems Security) at

Concordia University

Montréal, Québec, Canada

August 2021

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Zixiang Xian**

Entitled: **Generative Models Based on the Bounded Asymmetric Gaussian Distribution**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Information Systems Security)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

______________________________ Chair, Examiner
Dr. Arash Mohammadi

______________________________ Examiner
Dr. Mohsen Ghafouri

______________________________ Supervisor
Dr. Nizar Bouguila

______________________________ Co-supervisor
Dr. Manar Amayri

Approved by ______________________________
Abdessamad Ben Hamza, Chair
Department of Concordia Institute for Information Systems Engineering

2 August 2021 ______________________________
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Generative Models Based on the Bounded Asymmetric Gaussian Distribution

Zixiang Xian

The bounded asymmetric Gaussian mixture model (BAGMM) has proved that it generally performs better than the classical Gaussian mixture model. In this thesis, we investigate the learning of the BAGMM. Indeed, we propose an Expectation-Maximization (EM) algorithm to estimate the model parameters. A model selection criterion for BAGMM using minimum message length (MML) is proposed to determine the optimal number of clusters. The MML is shown to perform better than other model selection criteria.

In this thesis, we additionally propose an unsupervised feature selection framework using BAGMM to determine the structure of high dimensional data without knowing in advance the number of clusters nor the importance of the involved features. The validation for this framework involves several human-related recognition challenges, such as human activity categorization and human gender recognition.

Finally, we integrate the BAGMM into a hidden Markov model (HMM) framework, which uses BAGMM to model the emission probability distribution. The BAGMM-based HMM is evaluated with several real-world applications and compared with other Gaussian mixture-based HMMs.

Acknowledgments

I would like to express my sincere gratitude to my supervisor $Prof.NizarBouguila$ because of his academic research guidance throughout the study of my master's program. His endless patience and valuable advice help me get on occasions whenever I am stuck at research. I am grateful to him not only for what I learn from him but also for his influence on my entire life.

I also extend my gratefulness to my co-supervisor, $Dr.ManarAmayri$, for her technical advice. Also, I am very fortunate to work with $Dr.MuhammadAzam$, who always motivates me with his timely support and passion. I would not be able to complete this research program without his constructive suggestions. Besides, an excellent thank you from my heart to all my fellow lab mates.

Finally, I am deeply grateful to my parents and my wife for their unconditional support and immense encouragement throughout my study in Canada. Without their moral and financial support, I would not have been able to complete my studies alone.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Finite mixture models [1, 2] are widely applied in a wide range of applications for pattern recognition, statistical inference, data mining and information retrieval because of their sound mathematical basis as an unsupervised learning approach. Mixture models are used for data clustering and Gaussian mixture model (GMM) is a well-known Bayesian learning method widely used in many applications. However, Gaussian distribution is symmetric and sensitive to outliers. In real life, data can be asymmetric. Hence, in order to improve the robustness of existing learning approaches and improve their modeling capabilities for asymmetric data, the asymmetric Gaussian mixture model (AGMM) has been proposed in [3]. On the other hand, the mixture of generalized Gaussian distributions (GGD) [4, 5, 6, 7, 8, 9, 10, 11] was proposed to overcome the drawback of GMM's rigidity in terms of shape and has been applied to many real applications [12, 13, 14, 15, 16]. Nevertheless, data lie in a bounded support range in many real applications, whereas algorithms to model these datasets have an unbounded support range. So the bounded support Gaussian mixture model (BGMM) was proposed in [17, 18, 19] to better model real-world data. Several bounded support mixtures have been proposed so far to improve the modeling capabilities of these algorithms for clustering [20, 21, 22]. Bounded asymmetric Gaussian mixture model (BAGMM) has been proposed in [23] and successfully applied to several applications. Our work is based on the BAGMM, which has been proven to perform better than the AGMM in clustering tasks [23] due to its bounded support and non-symmetric nature.

1.1.1 Model Selection Criterion for BAGMM

The most general approach for parameter estimation in mixture models is based on maximizing the log-likelihood function efficiently through the Expectation-Maximization (EM) framework [24, 25, 26, 27] but an essential part of modeling is to determine the optimal number of clusters. In general, there are many ways to achieve this by either deterministic or stochastic ways. The general stochastic approach uses Markov Chain Monte Carlo (MCMC) methods to either implement the model selection criteria [28] or implement the fully Bayesian inference by approximating the posterior distribution to find the optimal number of clusters. In this thesis, we focus on deterministic approaches and propose a model selection criterion for BAGMM using minimum message length (MML) [4, 3]. We also validate our proposed MML on synthetic and real-world datasets by comparing it with other model selection criteria, including Akaike's information criterion (AIC) [29], the Schwarz's Bayesian information criterion (BIC) [30], Consistent AIC (CAIC) [31], minimum description length (MDL) [32], the mixture minimum description length (MMDL) [33] and the Laplace empirical criterion (LEC) [34].

1.1.2 Feature Selection Using BAGMM

Modern computer vision applications generate high-dimensional vectors which are challenging to model [35]. 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 computational complexity. Hence, irrelevant features 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 [36], 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, [37, 38, 39, 40]). This thesis investigates the effectiveness of feature selection using BAGMM (BAGMM-FS) in several human related recognition challenging tasks such as human action recognition and gender recognition. 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.

1.1.3 BAGMM Integration into the HMM framework

Statistical methods of Markov source, known as hidden Markov modeling, were initially introduced and studied in the late 1960s and early 1970s since Baum and his colleagues published a theory about estimating HMM parameters given a training observation sequence via the maximum likelihood (ML) method [41, 42, 43, 44]. The fundamental motivation behind the adoption of HMMs is to characterize real-world signals regarding the signal models, which may help us enhance signals by removing noise and transmission distortion, and to learn the details about the signal source without having to have the source available via simulations [45]. Besides, they have been proved to be very practical while dealing with non-observable data over a time interval to disclose the future values or reveal the latent variables. Although there are some research works that tend to improve the HMM structure by tuning the initialization step in the context of parameters setting [46, 47], the training process of HMM remains the identical regulated form via the Expectation-Maximization algorithm [48]. However, in most cases, the choice of emission probability distributions is less discussed and adopted by Gaussian mixture models (GMM) by default, often because of mathematical and practical convenience and strong assumption of a common pattern for real data [49]. However, this strong assumption is potentially insufficient to achieve the best modeling performance, as discussed in Section 1.1, while BAGMM can overcome the drawbacks of assuming symmetric unbounded data. For this reason, we propose to explore and evaluate the performance of HMM by adopting BAGMM as emission probability distribution and compare it with the Gaussian mixture model-based HMM (GMM-HMM) and other Gaussian-based generalizations. The details on the parameters learning process of the proposed model, including the parameters setting, i.e., the number of hidden states and mixture components, but also the performance, will be discussed in this thesis.

1.2 Contributions

The contributions of this thesis are as follows:

☞ **Model Selection Criterion for Multivariate Bounded Asymmetric Gaussian Mixture Model**

We propose model selection criterion for bounded support asymmetric Gaussian mixture model (BAGMM) using minimum message length (MML). The proposed approach is applied to several sets of synthetic data, real data and occupancy detection. The results of model selection and clustering are compared with other model selection criteria and asymmetric Gaussian mixture model (AGMM). This contribution has been accepted by the $29th$ European Signal Processing Conference, EUSIPCO 2021 [50].

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

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 evaluations involved several human-related recognition challenges. This work has been accepted by the IEEE 22^{nd} International Conference on Information Reuse and Integration for Data Science [51].

☞ **Bounded asymmetric Gaussian mixture-based hidden Markov models**

We first introduce a complete derivation of the equations for integrating the bounded asymmetric Gaussian mixture into the HMM framework and apply this innovative HMM framework to real-world applications while comparing it with traditional HMM and Gaussian mixture-derived HMMs. This research work has been accepted as a book chapter [52].

1.3 Thesis Overview

The rest of this thesis is organized as follows:

- Chapter 2 introduces the asymmetric Gaussian mixture model (AGMM) and bounded asymmetric Gaussian mixture model (BAGMM) briefly. Then the derivation of the model selection criterion for BAGMM using minimum message length (MML) is discussed in detail, as well as the experiments on synthetic datasets and real-world applications.

- Chapter 3 is devoted to applying an unsupervised feature selection framework based on the expectation-maximization (EM) algorithm using BAGMM to model high-dimensional data without knowing the number of clusters or the weights of involved features. Several challenging human-related image datasets are selected for validation of the proposed approach.

- Chapter 4 describes the first integration of BAGMM into the framework of hidden Markov models (HMM) using the EM approach. We evaluated the proposed innovative HMM with the tasks of occupancy estimation and human activity recognition (HAR) compared with other comparable Gaussian mixture-based HMMs.

- In conclusion, we summarize our work and contributions with some remarks for potential future research.

Chapter 2

Model Selection Criterion for Bounded Asymmetric Gaussian Mixture Model

This chapter proposes a model selection criterion for bounded support asymmetric Gaussian mixture model (BAGMM) using minimum message length (MML). The proposed approach is validated using synthetic data, real-world data, and occupancy detection applications. The proposed method is compared with other state-of-the-art model selection approaches. Moreover, the developed bounded mixture is compared with the asymmetric Gaussian mixture model (AGMM).

2.1 Mixture of Asymmetric Gaussian Distributions

The asymmetric Gaussian mixture model (AGMM) was designed to handle non-symmetric data sets [3, 40, 53]. Given a D-dimensional random variable $\mathbf{X} = [X_1, \ldots, X_D]$ that follows a M-component mixture of distributions, where the PDF associated with each component is the multidimensional asymmetric Gaussian distribution (AGD) [54, 55, 56, 57, 58, 59, 60]:

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

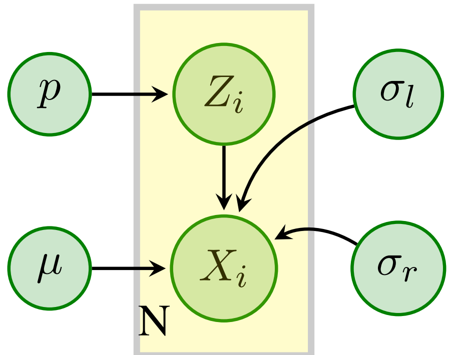

Figure 1: Graphical representation of the asymmetric Gaussian mixture model

where $\xi_m = (\vec{\mu}_m, \sigma_{\vec{l}_m}, \sigma_{\vec{r}_m})$ represents the parameters of AGD. Here, $\vec{\mu}_m = (\mu_{m1}, ..., \mu_{mD})$, $\sigma_{\vec{l}_m} = (\sigma_{l_{m1}}, ..., \sigma_{l_{mD}})$, and $\sigma_{\vec{r}_m} = (\sigma_{r_{m1}}, ..., \sigma_{r_{mD}})$ are the mean, left standard deviation and right standard deviation of the D-dimensional AGD, respectively. The parameters of AGMM are learnt via Expectation-Maximization (EM), and details are explained in [3, 40, 53]. Graphical representation of AGMM is displayed in Figure. 1, where X_i is one of N data instances with $i = 1, \ldots, N$. μ, σ_l and σ_r are the parameters of the distribution; p and Z_i are the mixing coefficient and posterior probability in the mixure model, which will be explained in Section. 2.2

2.2 Mixture of Bounded Asymmetric Gaussian Distributions

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

$$p(\vec{X}|\Theta) = \sum_{j=1}^{K} p(\vec{X}|\xi_j) p_j \tag{2}$$

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:

$$p(\vec{X}|\xi_j) = \frac{f(\vec{X}|\xi_j)\mathrm{H}(\vec{X}|\Omega_j)}{\int_{\partial_j} f(\vec{\mathrm{u}}|\xi_j) d\mathrm{u}}, \text{where } \mathrm{H}(\vec{X}|\Omega_j) = \begin{cases} 1 & \text{if} \quad \vec{X} \in \partial_j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$f(\vec{X}|\xi_j) = \prod_{d=1}^{D} \frac{2}{\sqrt{2\pi}(\sigma_{l_{jd}} + \sigma_{r_{jd}})} \times \begin{cases} \exp\left[-\frac{(\mathrm{X}_d-\mu_{jd})^2}{2\sigma_{l_{jd}}^2}\right] & \text{if} \quad \mathrm{X}_d < \mu_{jd} \\ \exp\left[-\frac{(\mathrm{X}_d-\mu_{jd})^2}{2\sigma_{r_{jd}}^2}\right] & \text{if} \quad \mathrm{X}_d \geq \mu_{jd} \end{cases} \tag{4}$$

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. (3) 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:

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

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

$$Z_{ij} = p(j|\vec{X}_i) = \frac{p(\vec{X}_i|\xi_j)p_j}{\sum_{j=1}^{K} p(\vec{X}_i|\xi_j)p_j} \quad \text{and} \quad \mathcal{Z} = \{\vec{Z}_1, ..., \vec{Z}_N\}. \tag{6}$$

2.3 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:

$$\text{Mess Len}(K) \simeq -\log(p(\Theta_K)) - \mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{1}{2}\log|F(\Theta_K)| + \frac{N_p}{2}(1 + \log(K_{N_p})) \tag{7}$$

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. K_{N_p} is the optimal quantization lattice constant, which can be approximated it by $\frac{1}{12}$. The estimation of number of classes is carried out by finding minimum with respect to Θ of message length [4, 3]. The derivation of $p(\Theta_K)$ and $|F(\Theta_K)|$ is given in following subsections.

2.3.1 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 :

$$p(\Theta) = p(\boldsymbol{\pi})p(\boldsymbol{\mu})p(\boldsymbol{\sigma}_l)\,p(\boldsymbol{\sigma}_r) \tag{8}$$

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:

$$p(\boldsymbol{\pi}) = \frac{\Gamma(\sum_{j=1}^{K} \eta_j)}{\sum_{j=1}^{K} \Gamma(\eta_j)} \sum_{j=1}^{K} {p_j}^{\eta_j - 1} \tag{9}$$

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:

$$p(\mu_{jd}) = \frac{1}{\sigma_{ld} + \sigma_{rd}} \implies p(\vec{\mu}_j) = \prod_{d=1}^{D} \frac{1}{\sigma_{ld} + \sigma_{rd}} \tag{10}$$

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{K} \prod_{d=1}^{D} \frac{1}{\sigma_{ld} + \sigma_{rd}} = \prod_{d=1}^{D} \frac{1}{(\sigma_{ld} + \sigma_{rd})^K} \tag{11}$$

For the parameter σ_l and σ_r, we have:

$$p(\boldsymbol{\sigma}_l) = \prod_{j=1}^{K} p(\vec{\sigma}_{l_j}), \quad p(\boldsymbol{\sigma}_r) = \prod_{j=1}^{K} p(\vec{\sigma}_{r_j}) \tag{12}$$

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:

$$p(\sigma_{l_{jd}}) = \frac{1}{\sigma_{l_d}}, \quad p(\sigma_{r_{jd}}) = \frac{1}{\sigma_{r_d}} \tag{13}$$

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

$$p(\vec{\sigma}_{l_j}) = \prod_{d=1}^{D} \frac{1}{\sigma_{l_d}}, \quad p(\vec{\sigma}_{r_j}) = \prod_{d=1}^{D} \frac{1}{\sigma_{r_d}} \tag{14}$$

From Eqs. (12 & 14), we obtain:

$$p(\boldsymbol{\sigma}_l) = \prod_{j=1}^{K}\prod_{d=1}^{D} \frac{1}{\sigma_{l_d}} = \prod_{d=1}^{D} \frac{1}{{\sigma_{l_d}}^K}, \quad p(\boldsymbol{\sigma}_r) = \prod_{j=1}^{K}\prod_{d=1}^{D} \frac{1}{\sigma_{r_d}} = \prod_{d=1}^{D} \frac{1}{{\sigma_{r_d}}^K} \tag{15}$$

Finally, by replacing the priors of parameters in Eq. (8) by Eqs. (11 & 15), we get:

$$p(\Theta) = (M-1)! \prod_{d=1}^{D} \frac{1}{{\sigma_{l_d}}^K {\sigma_{r_d}}^K (\sigma_{ld} + \sigma_{rd})^K} \tag{16}$$

2.3.2 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. (17).

$$|F(\Theta)| = |F(\boldsymbol{\pi})|\,|F(\boldsymbol{\mu})|\,|F(\boldsymbol{\sigma}_l)|\,|F(\boldsymbol{\sigma}_r)| \tag{17}$$

$$|F(\boldsymbol{\pi})| = \frac{N^{K-1}}{\sum_{j=1}^{K} p_j}, \quad F(\vec{\mu}_j)_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial \mu_{jk_1} \partial \mu_{jk_2}} \tag{18}$$

$$F(\vec{\sigma}_{l_j})_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial \sigma_{l_{jk_1}} \partial \sigma_{l_{jk_2}}}, \quad F(\vec{\sigma}_{r_j})_{k_1,k_2} = \frac{\partial^2 \mathcal{L}(\Theta, Z, \mathcal{X}_j)}{\partial \sigma_{r_{jk_1}} \partial \sigma_{r_{jk_2}}} \tag{19}$$

$$\begin{aligned}
|F(\boldsymbol{\mu})| = \prod_{j}^{K} \prod_{d=1}^{D} \Bigg| & \sum_{i=l,\mathrm{X}_{id}<\mu_{jd}}^{l+n_j-1} \left[\frac{-1}{\sigma_{l_{jd}}^2}\right] + \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \left[\frac{-1}{\sigma_{r_{jd}}^2}\right] \\
& - \sum_{i=l,\mathrm{x}_d<\mu_{jd}}^{l+n_j-1} \frac{1}{\sigma_{l_{jd}}^4} \times \Bigg\{ - \frac{\left(\frac{1}{M}\sum_{m=1}^{M} (\mathrm{l}_{m_{jd}} - \mu_{jd}) \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M} \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)\right)^2} \\
& + \frac{\frac{1}{M}\sum_{m=1}^{M} \left\{(\mathrm{l}_{m_{jd}} - \mu_{jd})^2 - 1\right\} \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M} \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j)} \Bigg\} \\
& - \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \frac{1}{\sigma_{r_{jd}}^4} \times \Bigg\{ - \frac{\left(\frac{1}{M}\sum_{m=1}^{M} (\mathrm{r}_{m_{jd}} - \mu_{jd}) \mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M} \mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)\right)^2} \\
& + \frac{\frac{1}{M}\sum_{m=1}^{M} \left\{(\mathrm{r}_{m_{jd}} - \mu_{jd})^2 - 1\right\} \mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M} \mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)} \Bigg\} \Bigg|
\end{aligned} \tag{20}$$

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

$$
\begin{aligned}
|F(\boldsymbol{\sigma}_r)| = \prod_{j}^{K}\prod_{d=1}^{D}\Bigg| -3 \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{r_{jd}}^4}\right) \\
- \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \left(\frac{-2}{\sigma_{r_{jd}}^3(\sigma_{l_{jd}}+\sigma_{r_{jd}})}\right) \left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\} \\
- \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \frac{1}{\sigma_{r_{jd}}^6} \left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^4\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\} \\
- \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \frac{-3}{\sigma_{r_{jd}}^4} \left\{\frac{\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)}\right\} \\
- \sum_{i=l,\mathrm{X}_{id}\geq\mu_{jd}}^{l+n_j-1} \frac{1}{\sigma_{r_{jd}}^6} \left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j)\right)^2}{(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j))^2}\right\}\Bigg|
\end{aligned} \tag{22}
$$

where $\mathrm{l}_{m_{jd}}$ is a set of random variables drawn from the asymmetric Gaussian distribution (AGD) with the constraint, 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, $\mathrm{r}_{m_{jd}}$ are the random variables drawn from the AGD with constraint, u $\geq \mu_{jd}$ for the particular component j of the mixture model.

2.3.3 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. (7). 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

2.4 Experimental Results

2.4.1 Comparison with other model selection criteria

The model selection criteria selected to compare with MML, include MDL [32], AIC [29], Bayesian inference criterion (BIC) [30], consistent AIC (CAIC) [31], mixture MDL (MMDL) [33], MML_{like} [61], LEC [62, 63]. The details of these algorithms is given in [64].

The following equation is the general form of deterministic model selection criteria.

$$C(\hat{\Theta}(K), K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) + f(K) \tag{23}$$

where $f(K)$ is an increasing function which penalizes higher values of K and optimal number of clusters in a mixture can be calculated as follows:

$$\hat{K} = arg\, min\{C(\hat{\Theta}(K), K), K = K_{min}, ..., K_{max}\} \tag{24}$$

Although all model selection criteria share this common point and utilize log-likelihood, they have different concepts, and their equations are described in the following.

$$MDL(K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{N_p}{2}\log(N) \tag{25}$$

where N_p is the number of free mixture parameters and computed as $K(3D+1)-1$ in this model.

$$AIC(K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{N_p}{2} \tag{26}$$

$$BIC(K) = -2\mathcal{L}(\Theta_K, Z, \mathcal{X}) + N_p\log(N) \tag{27}$$

$$CAIC(K) = -2\mathcal{L}(\Theta_K, Z, \mathcal{X}) + N_p(1+\log(N)) \tag{28}$$

$$MMDL(K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{1}{2} N_p \log(N) + \frac{c}{2} \sum_{j=1}^{K} \log(p_j) \tag{29}$$

where c is the number of free parameters for each mixture component and computed as $3D+1$ in this model.

$$MML_{Like}(K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) + \frac{K}{2} \log\left(\frac{N}{12}\right) + \frac{c}{2} \sum_{j=1}^{K} \log\left(N\frac{p_j}{12}\right) + \frac{N_p}{2} \tag{30}$$

For model selection via LEC, we have:

$$LEC(K) = -\mathcal{L}(\Theta_K, Z, \mathcal{X}) - \log(P(\Theta_K)) - \frac{1}{2} N_p \log(2\pi) + \frac{1}{2} \log(|F(\Theta_K)|) \tag{31}$$

2.4.2 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 1 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 1, 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 2 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.

Table 1: 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 2: 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

2.4.3 Real Datasets

We have adopted 10 standard multidimensional datasets to validate the proposed model with real datasets, which 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 [65]. 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 Table 3. 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.

Table 3: 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**

2.4.4 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 [66] is composed of 9752 instances, 5 dimensions and 2 clusters, as shown in the Table 4. 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 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 2 and Figure 3 displays the results of different model selection criteria for BAGMM and AGMM, respectively. 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 4: 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%**

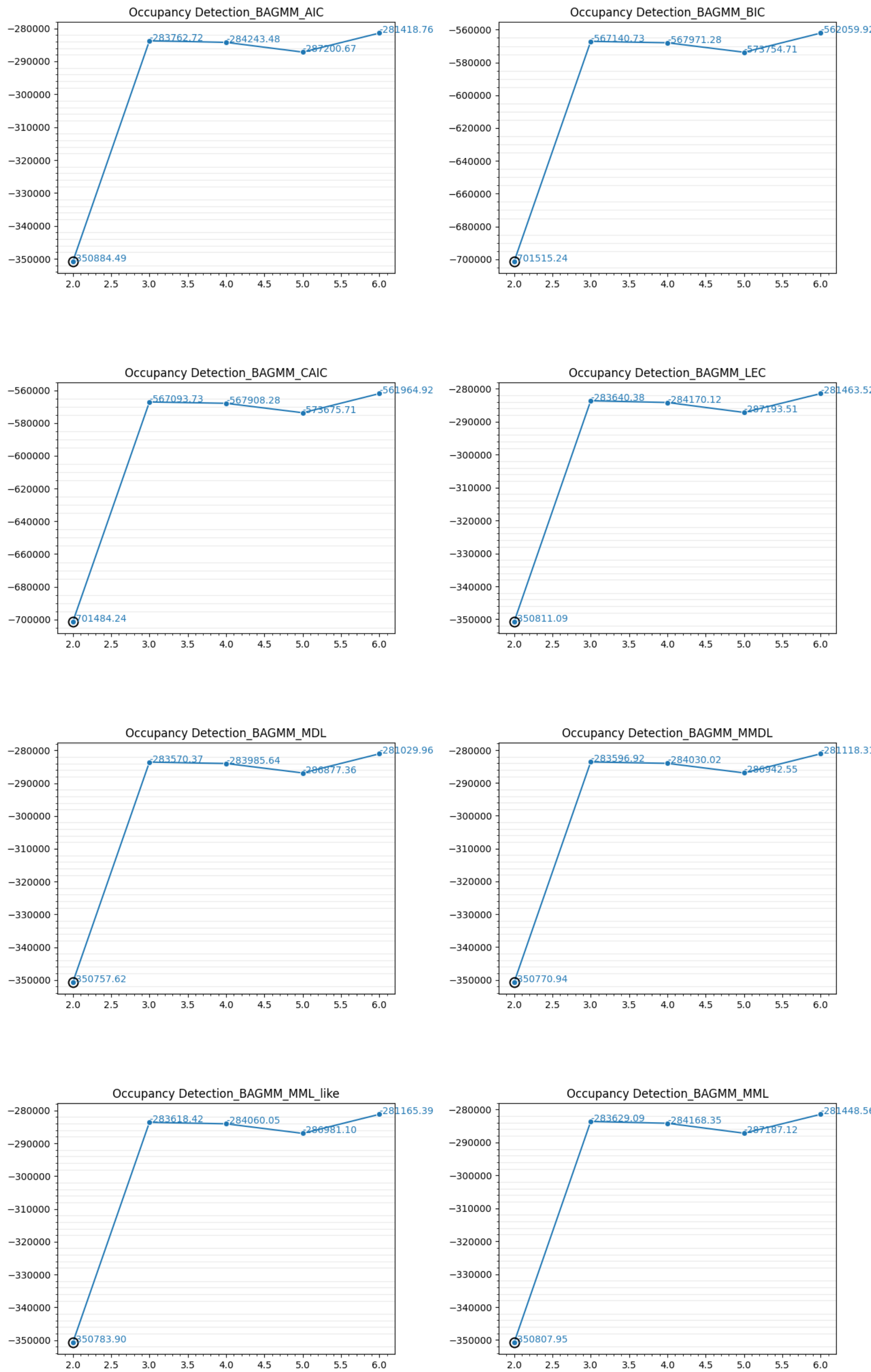

Figure 2: Different Model Selection Criteria for Occupancy Dataset for **BAGMM**

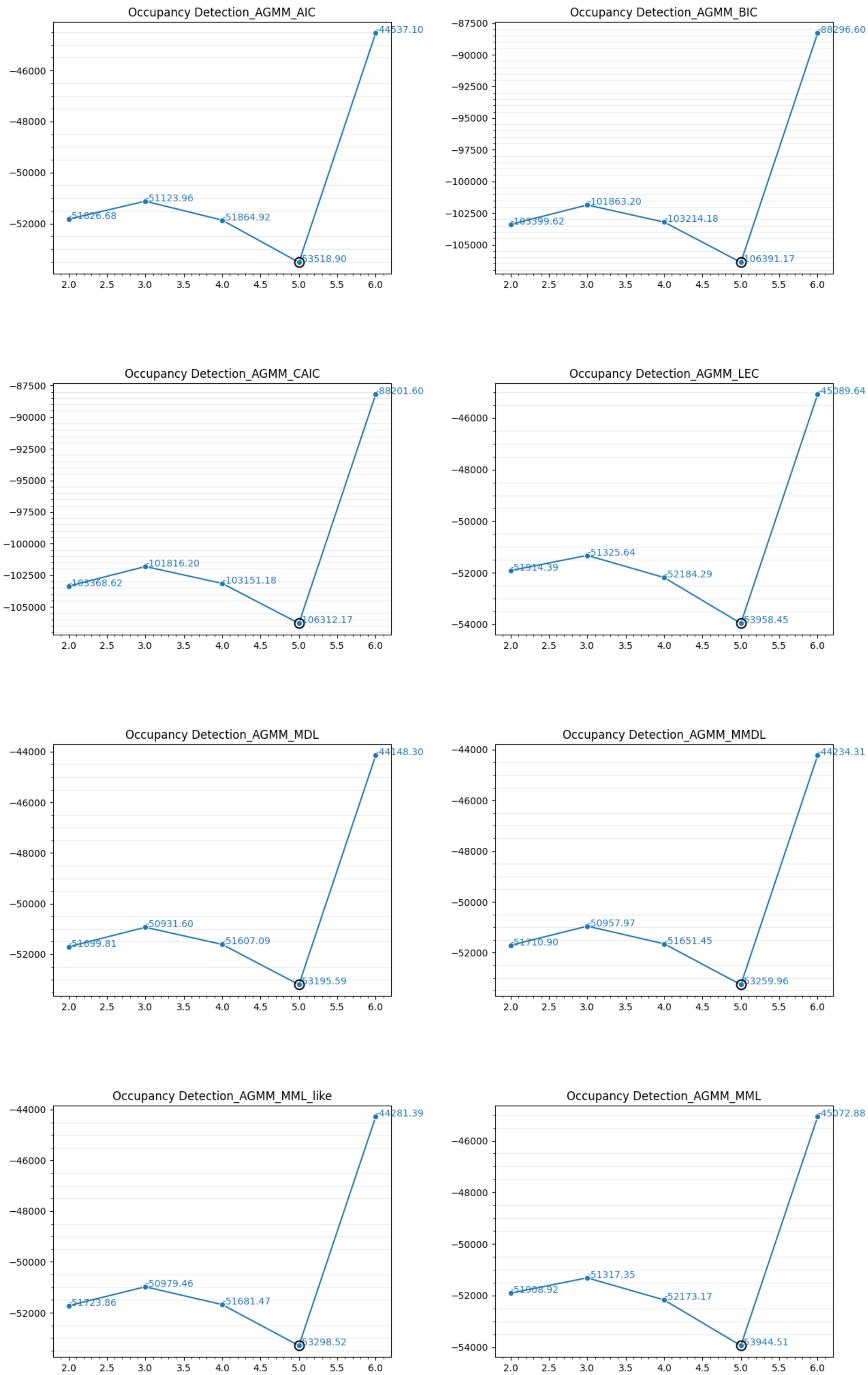

Figure 3: Different Model Selection Criteria for Occupancy Dataset for **AGMM**

Chapter 3

Feature Selection Using Bounded Asymmetric Gaussian Mixtures: Application to Human Action and Gender Recognition

In the previous chapter, we have proposed a model selection criterion for bounded support asymmetric Gaussian mixture model (BAGMM) using minimum message length (MML). In this chapter, we propose an unsupervised feature selection framework using the bounded asymmetric Gaussian mixture model (BAGMM-FS) to determine the structure of high dimensional data without prior knowledge of the number of clusters and the importance of the involved features. In many cases, irrelevant features can compromise the effectiveness of clustering and increase computational complexity, so irrelevant features should be given small weights or even discarded. Selecting a relevant feature space is generally known as feature selection or also called subset selection.

3.1 Proposed Model

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:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} \sum_{j=1}^{K} p(\vec{X}_i|\xi_j)p_j \tag{32}$$

where $p(\vec{X}_i|\xi_j)$ is the PDF of BAGD defined in Section. 2.2 of Eq. (3).

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:

$$p(\mathcal{X}, \mathcal{Z}|\Theta) = \prod_{i=1}^{N} \prod_{j=1}^{K} \left(p(\vec{X}_i|\xi_j)p_j\right)^{Z_{ij}} \tag{33}$$

We can get the complete data log-likelihood by taking the logarithm of Eq. (33) as follows.

$$\log p(\mathcal{X}, Z \mid \Theta) = \sum_{i=1}^{N} \sum_{j=1}^{K} Z_{ij} \log \left[p\left(\vec{X}_i \mid \xi_j\right) p_j\right] \tag{34}$$

where $\mathcal{Z} = \left\{\vec{Z}_1, \ldots, \vec{Z}_N\right\}$. According to Eq. (34), all the D features in the model have the same weight which can not describe well real-world data since some features may be irrelevant for 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 [67] in the case of the finite Gaussian mixture. Then, the resulting model can be rewritten as:

$$p(\vec{X}_i \mid \Theta, \vec{\lambda}, \vec{\varphi}) = \sum_{j=1}^{K} p_j \prod_{d=1}^{D} p\left(X_d \mid \xi_{jd}\right)^{\varphi_d} p\left(X_d \mid \lambda_d\right)^{1-\varphi_d} \tag{35}$$

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 [67], we can obtain:

$$p\left(\vec{X}_i \mid \Theta_K\right)=\sum_{j=1}^{K} p_j \prod_{d=1}^{D}\left[\omega_d p\left(X_d \mid \xi_{jd}\right)+\left(1-\omega_d\right) p\left(X_d \mid \lambda_d\right)\right] \tag{36}$$

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 $0 \leq \omega_d \leq 1$, $d=1,\ldots,D$.

3.2 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:

$$\hat{Z}_{ij}=\frac{p_j \prod_{d=1}^{D} \phi_{ijd}}{\sum_{j=1}^{K} p_j \prod_{d=1}^{D} \phi_{ijd}} \tag{37}$$

where

$$\phi_{ijd}=\omega_d p\left(X_{id} \mid \xi_{jd}\right)+\left(1-\omega_d\right) p\left(X_{id} \mid \lambda_d\right) \tag{38}$$

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

$$\begin{aligned}\mathcal{L}(\mathcal{X},\mathcal{Z} \mid \Theta)&=\sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\log\left(p\left(\vec{X}_i \mid \Theta_K\right)\right)\\&=\sum_{i=1}^{N}\sum_{j=1}^{K} Z_{ij}\left\{\log p_j+\log\left[\omega_d p\left(\vec{X}_i \mid \xi_j\right)+\left(1-\omega_d\right) p\left(\vec{X}_i \mid \lambda\right)\right]\right\}\end{aligned} \tag{39}$$

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:

$$p_j^{new}=\frac{\sum_{i=1}^{N} h\left(j \mid \vec{X}_i,\Theta_M\right)}{N} \tag{40}$$

$$\mu_{jd}^{new}=\frac{\sum_{i=1}^{N}\frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i,\Theta_M\right)\left\{\mathbf{X}_{id}-\frac{\int_{\partial_j} f(\mathbf{u}|\xi_j)\left(\mathbf{u}-\mu_{jd}\right)d\mathbf{u}}{\int_{\partial_j} f(\mathbf{u}|\xi_j)d\mathbf{u}}\right\}}{\sum_{i=1}^{N}\frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i,\Theta_M\right)} \tag{41}$$

Note that in Eq. (41), 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:

$$\int_{\partial_j} f(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}(s_{m_{jd}}-\mu_{jd})\mathrm{H}(s_{m_{jd}}|\Omega_j) \tag{42}$$

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. (41) can be approximated as:

$$\int_{\partial_j} f(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(s_{m_{jd}}|\Omega_j) \tag{43}$$

Thus, μ_{jd}^{new} can be written as:

$$\mu_{jd}^{new} = \frac{\sum_{i=1}^{N}\frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i,\Theta_M\right)\left\{\mathrm{X}_{id} - \frac{\sum_{m=1}^{M}(s_{m_{jd}}-\mu_{jd})\mathrm{H}(s_{m_{jd}}|\Omega_j)}{\sum_{m=1}^{M}\mathrm{H}(s_{m_{jd}}|\Omega_j)}\right\}}{\sum_{i=1}^{N}\frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}} h\left(j \mid \vec{X}_i,\Theta_M\right)} \tag{44}$$

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 :

$$\sigma_{l_{jd}}^{new} = \sigma_{l_{jd}}^{old} - \left[\left(\frac{\partial^2\mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma_{l_{jd}}^2}\right)^{-1}\left(\frac{\partial\mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma_{l_{jd}}}\right)\right] \tag{45}$$

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

$$\begin{aligned}\frac{\partial\mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma_{l_{jd}}} &= \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N}\frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}}\times h\left(j\mid\vec{X}_i,\theta_M\right)\left(\frac{\left(\mathrm{X}_{id}-\mu_{jd}\right)^2}{\sigma_{l_{jd}}^3}\right)\\ &-\sum_{\mathrm{X}_{id}<\mu_{jd}}^{N}\frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}\times\sigma_{l_{jd}}^3} h\left(j\mid\vec{X}_i,\theta_M\right)\times\left\{\frac{\int_{\partial_j}\mathrm{g}_1\left(\mathrm{u}\mid\xi_j\right)\left(\mathrm{u}-\mu_{jd}\right)^2d\mathrm{u}}{\int_{\partial_j}\mathrm{g}_1\left(\mathrm{u}\mid\xi_j\right)d\mathrm{u}}\right\}\end{aligned} \tag{46}$$

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:

$$\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 du \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{m_{jd}}-\mu_{jd})^2 \mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j) \tag{47}$$

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. (46) can be approximated as:

$$\int_{\partial_j} \mathrm{g}_1(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{m_{jd}}|\Omega_j) \tag{48}$$

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:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial \sigma_{l_{jd}}^2} &= -3\sum_{X_{id}<\mu_{jd}}^{N}\gamma_{ij}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{l_{jd}}^4}\right) \\
&- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N}\gamma_{ij}\left(\frac{-2}{\sigma_{l_{jd}}^3\left(\sigma_{l_{jd}}+\sigma_{r_{jd}}\right)}\right)\times\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{l}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}\right\} \\
&- \sum_{X_{id}<\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma_{l_{jd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{l}_{m_{jd}}-\mu_{jd}\right)^4\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}\right\} \\
&- \sum_{\mathrm{X}_{id}<\mu_{jd}}^{N}\frac{-3\gamma_{ij}}{\sigma_{l_{jd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{l}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)}\right\} \\
&- \sum_{X_{id}<\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma_{l_{jd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{l}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{l}_{m_{jd}}\mid\Omega_j\right)\right)^2}\right\}
\end{aligned} \tag{49}$$

where

$$\gamma_{ij} = \frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}}Z_{ij} \tag{50}$$

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

$$\sigma_{r_{jd}}^{new} = \sigma_{r_{jd}}^{old} - \left[\left(\frac{\partial^2\mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma_{r_{jd}}^2}\right)^{-1}\left(\frac{\partial\mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial\sigma_{r_{jd}}}\right)\right] \tag{51}$$

where

$$\frac{\partial \mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial \sigma_{r_{jd}}} = \sum_{i=1,\mathrm{X}_{id}\geq\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}} \times h\left(j\mid\vec{X}_i,\theta_M\right)\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{r_{jd}}^3}\right)$$
$$- \sum_{i=1,\mathrm{X}_{id}\geq\mu_{jd}}^{N} \frac{\omega_d p\left(X_{id}\mid\xi_{jd}\right)}{\phi_{ijd}\times\sigma_{l_{jd}}^3} h\left(j\mid\vec{X}_i,\theta_M\right)\times\left\{\frac{\int_{\partial_j}\ \mathrm{g}_2\left(\mathrm{u}\mid\xi_j\right)(\mathrm{u}-\mu_{jd})^2\,d\mathrm{u}}{\int_{\partial_j}\ \mathrm{g}_2\left(\mathrm{u}\mid\xi_j\right)d\mathrm{u}}\right\} \tag{52}$$

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:

$$\int_{\partial_j}\mathrm{g}_2(\mathrm{u}|\xi_j)(\mathrm{u}-\mu_{jd})^2 d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{m_{jd}}-\mu_{jd})^2\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j) \tag{53}$$

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. (52) can be approximated as:

$$\int_{\partial_j}\mathrm{g}_2(\mathrm{u}|\xi_j)d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{m_{jd}}|\Omega_j) \tag{54}$$

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

$$\frac{\partial^2 \mathcal{L}(\mathcal{X},\mathcal{Z}\mid\Theta)}{\partial \sigma_{r_{jd}}^2} = -3\sum_{X_{id}\geq\mu_{jd}}^{N}\gamma_{ij}\left(\frac{(\mathrm{X}_{id}-\mu_{jd})^2}{\sigma_{r_{jd}}^4}\right)$$
$$-\sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\gamma_{ij}\left(\frac{-2}{\sigma_{r_{jd}}^3\left(\sigma_{r_{jd}}+\sigma_{r_{jd}}\right)}\right)\times\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\}$$
$$-\sum_{X_{id}\geq\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma_{r_{jd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^4\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\}$$
$$-\sum_{\mathrm{X}_{id}\geq\mu_{jd}}^{N}\frac{-3\gamma_{ij}}{\sigma_{r_{jd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)}\right\}$$
$$-\sum_{X_{id}\geq\mu_{jd}}^{N}\frac{\gamma_{ij}}{\sigma_{r_{jd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathbf{r}_{m_{jd}}-\mu_{jd}\right)^2\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathbf{r}_{m_{jd}}\mid\Omega_j\right)\right)^2}\right\} \tag{55}$$

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

$$\eta_d^{new} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i, \theta_M\right)\right]X_{id}}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i, \theta_M\right)} \tag{56}$$

$$\delta_d^{2^{\text{new}}} = \frac{\sum_{i=1}^{N}\left[\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i, \theta_M\right)\right](X_{id}-\eta_d)^2}{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{(1-\omega_d)p(X_{id}|\lambda_d)}{\phi_{ijd}}h\left(j \mid \vec{X}_i, \theta_M\right)} \tag{57}$$

$$\omega_d^{\text{new}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{\omega_d p\left(X_{id}|\xi_{jd}\right)}{\phi_{ijd}}h\left(j \mid \vec{X}_i, \theta_M\right)}{N} \tag{58}$$

3.2.1 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

$$\text{MessLens} \approx -\log p(\Theta_M) + \frac{c}{2}(1+\log(K_c)) + \frac{1}{2}\log|I(\Theta_M)| - \log p(\mathcal{X} \mid \Theta_M) \tag{59}$$

where $p(\Theta_M)$ is prior distribution, $I(\Theta_M)$ denotes the Fisher information matrix, $\log p(\mathcal{X} \mid \Theta_M)$ is log-likelihood. Here the constant value c represents the total number of free parameters, which is equal $M + D + 3DM + 2D$, K_c is the optimal quantization lattice constant, which can be approximated it by $\frac{1}{12}$, $|I(\Theta_M)|$ 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 $p(\Theta_M)$ and $I(\Theta_M)$. 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:

$$\begin{aligned}\text{MessLens} \approx & \frac{c}{2}(1+\log(K_c)) + \frac{c}{2}(\log N) + \frac{3M}{2}\sum_{d=1}^{D}\log\omega_d \\ & + \frac{3D}{2}\sum_{j=1}^{M}\log p_j + \sum_{d=1}^{D}\log(1-\omega_d) - \log p(\mathcal{X} \mid \theta_M)\end{aligned} \tag{60}$$

The minimization of the previous equation gives the following:

$$p_j^* = \frac{\max\left(\sum_{i=1}^{N} h\left(j \mid \vec{X}_i, \Theta_M\right) - \frac{3D}{2}, 0\right)}{\sum_{j=1}^{M} \max\left(\sum_{i=1}^{N} h\left(j \mid \vec{X}_i, \Theta_M\right) - \frac{3D}{2}, 0\right)} \tag{61}$$

$$\omega_d^* = \frac{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} a_{ijd} - \frac{3M}{2}, 0\right)}{\max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} U_{ijd} - \frac{3M}{2}, 0\right) + \max\left(\sum_{i=1}^{N}\sum_{j=1}^{M} V_{ijd} - 1, 0\right)} \tag{62}$$

where

$$U_{ijd} = h\left(j \mid \vec{X}_i, \Theta_M\right) \frac{\omega_d p\left(X_{id} \mid \xi_{jd}\right)}{\phi_{ijd}} \tag{63}$$

$$V_{ijd} = h\left(j \mid \vec{X}_i, \Theta_M\right) \frac{\left(1-\omega_d\right) p\left(X_{id} \mid \lambda_d\right)}{\phi_{ijd}} \tag{64}$$

The complete learning of BAGGM-FS is given in Algorithm 2, 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.

3.3 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 [68] which indicates the overlapping clusters with the range from -1 to 1, and 1 is the best

Algorithm 2 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. (37).
    {[**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. (40, 44, 45, & 51).
    end for
    Estimate $\vec{\eta}$, $\vec{\delta}$ & $\vec{\omega}$ using Eqs. (56, 57, & 58).
    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. (60)
end for

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 [69], which indicates good clustering when it is low and poor clustering when it is high.

3.3.1 Human Activity Categorization

Human activity categorization (HAR) has received a lot of research attention in the last decade [70, 71]. 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 [72]. 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 5, 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

Table 5: 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

MML for BAGMM-FS with the MML for BAGMM, AGMM-FS, and AGMM in Fig. 4. 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.

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 6, 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

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

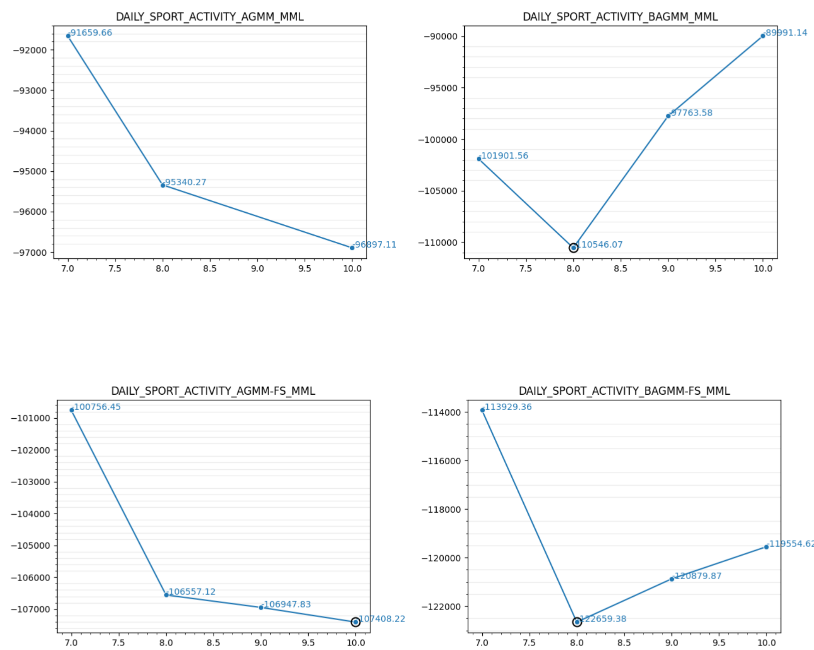

Figure 4: MML for the activities clustering application using different mixture models.

considered clustering metrics.

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

Table 6: Clustering of the sitting activities of the 8 subjects using different mixture models.

3.3.2 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 [73]). However, it's a complex task for machines because of people's variation status at different light intensities, such as different postures, angles, etc [74]. 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] [75, 76, 77, 78].

Figure 5: Samples images from PARSE-27k dataset.

Fig. 5 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 64×128 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

[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/

Figure 6: Samples images from CUHK sub-dataset in PETA dataset

resolutions of 80×160, a high camera angle, and a varying viewpoint. Fig. 6 shows sample images from the PETA dataset.

In order to describe the images, we have considered bag of visual words (BOVW) [79]. The basic idea is to extract local features for each image using scale invariant feature transform (SIFT) [80]. 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.

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 7 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 7 and we can see again that our model has an excellent performance as compared to the other models.

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	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	HR dataset	0.506	1	**70.61%**	73.89%	70.61%	69.56%	0.125	0.116
BAGMM	HR dataset	0.504	1	62.77%	78.66%	62.77%	56.79%	0.110	0.007
AGMM-FS	HR dataset	0.571	16	50.00%	25.00%	50%	33.33%	N/A	0.693
AGMM	HR dataset	4.003	300	50.00%	25.00%	50%	33.33%	N/A	0.693
BGGMM	HR dataset	48.073	8	61.98%	62.43%	61.98%	61.62%	0.097	0.009
GGMM	HR dataset	0.121	3	50.00%	25.00%	50%	33.33%	N/A	0.693
GMM	HR dataset	0.038	1	50.00%	25.00%	50%	33.33%	N/A	0.693

Table 7: Gender recognition results

We have also considered a challenging dataset called Human attribute (HR) dataset [77, 78]. 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.

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 7. 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.

Figure 7: Samples images from human attribute dataset.

Chapter 4

Bounded asymmetric Gaussian mixture-based hidden Markov models

In the previous chapter, we have proposed an unsupervised feature selection framework using the bounded asymmetric Gaussian mixture model (BAGMM-FS) and validated it on several human-related recognition challenges to prove its effectiveness. In this chapter, we integrate the bounded asymmetric Gaussian mixture model into a hidden Markov model (HMM) framework.

4.1 Introduction

In most cases, the choice of emission probability distributions in HMM is less discussed and considered as Gaussian mixture models (GMM) by default, often because of mathematical and practical convenience and strong assumption of a common pattern for the data [49]. In this Chapter, we propose to explore and evaluate the performance of HMM by adopting BAGMM as emission probability distribution.

Given a D-dimensional random variable $\mathbf{X} = [X_1, \dots, X_D]$, the bounded asymmetric Gaussian distribution (BAGD) for the vector $\vec{X}$ can be defined as:

$$p(\mathbf{X}|\xi_m) = \frac{f(\mathbf{X}|\xi_m)\mathrm{H}(\mathbf{X}|\Omega_m)}{\int_{\partial_m} f(\vec{\mathrm{u}}|\xi_m)d\mathrm{u}}, \text{where } \mathrm{H}(\mathbf{X}|\Omega_m) = \begin{cases} 1 & \text{if } \quad \vec{X} \in \partial_m \\ 0 & \text{otherwise} \end{cases} \tag{65}$$

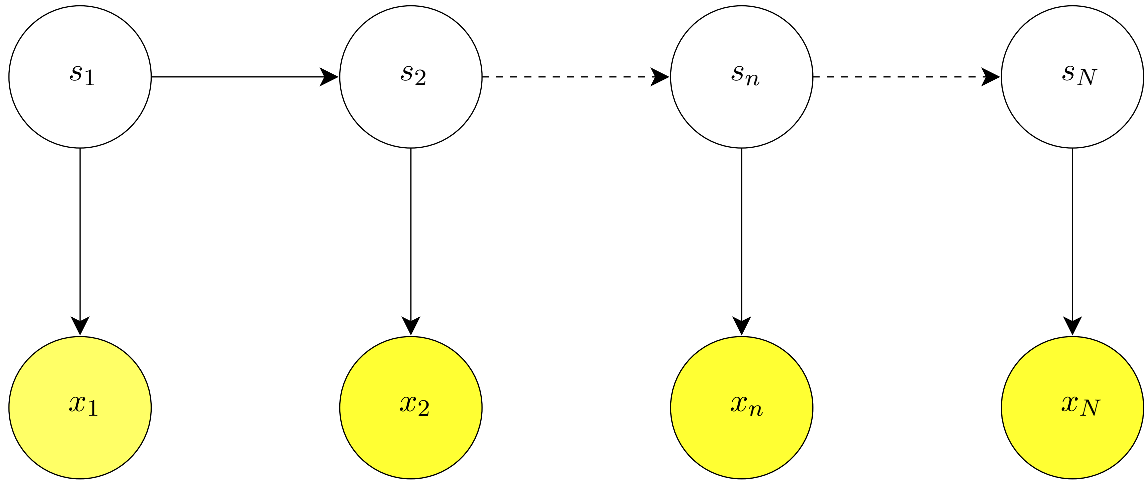

Figure 8: Graphical representation for HMM

where $f(\vec{X}|\xi_m)$ is the PDF of the asymmetric Gaussian distribution (AGD) defined in Section. 2.1 of Eq. (1), the term $\int_{\partial_m} f(\vec{\mathrm{u}}|\xi_m)d\mathrm{u}$ in Eq. (65) is the normalized constant that shows the share of $f(\vec{X}|\xi_m)$ which belongs to the support region ∂.

4.2 Hidden Markov Model

For many real-world applications, such as occupancy estimation in buildings, we wish to predict the following number of people in a time series given sequences of the previous values. It's impractical to consider a general dependence of future observations on all previous values. Therefore, the HMM assumes that the future predictions are dependent of the most recent observations only. Moreover, the HMM is a specific instance of the state space model that the latent variables are discrete. The latent variable, which is the state of this hidden process, satisfies the Markov property; that is, given the value of s_{n-1}; the current state s_n is independent of all the states prior to the time $n-1$. $X = [x_1, \ldots, x_N]$ represents the observed variables and $S = [s_1, \ldots, s_n]$ is the hidden state. A hidden Markov model is governed by a set of parameters, such as the set of state transitions and emission probability. There are three main tasks for HMM-based modeling; first is to optimize those parameters for the model given training data; second is scoring that calculates the joint probability of a sequence given the model; third is decoding that finds the optimal series of hidden states.

According to [81], given time series observations $X = [x_1, \ldots, x_n, \ldots, x_N]$ generated by hidden states $S = [s_1, \ldots, s_n, \ldots, s_N]$; $s_k \in [1, K]$ where K is the number of the hidden states, we define the transition probability matrix as A: $A_{jk} = p(s_{nk} = 1 \mid s_{n-1,k} = 1)$. They should satisfy $0 \leq A_{jk} \leq 1$ with $\sum_k A_{jk} = 1$, because they are probabilities.

$P(x_m \mid \Lambda)$ is known as emission probability, where Λ is a set of parameters governing the distribution if x is continuous. Note that $P(x_m \mid \Lambda)$ will be an emission probability matrix if x is discrete. The joint probability distribution over both hidden states and observed variables is then given by:

$$p(\mathbf{X}, \mathbf{S} \mid \mathbf{\Theta}) = p(\mathbf{s}_1 \mid \boldsymbol{\pi}) \left[\prod_{n=2}^{N} p(\mathbf{s}_n \mid \mathbf{s}_{n-1}, \mathbf{A}) \right] \prod_{m=1}^{N} p(\tilde{\mathbf{x}}_m \mid \Lambda) \tag{66}$$

where $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_N}]$, $\mathbf{S} = [\mathbf{s_1}, \ldots, \mathbf{s_N}]$, and $\mathbf{\Theta} = \{\boldsymbol{\pi}, \mathbf{A}, \mathbf{\Lambda}\}$ defines the set of parameters of HMM. Indeed, there are a wide range of choices for emission distribution that include Gaussian distribution and mixture models such as Gaussian mixture model (GMM). It's worth mentioning that the emission distributions are often taken as Gaussian mixtures for most continuous observations cases [81, 82, 83, 84].

The parameters learning task is crucial for HMM. In this paper, we focus on the maximum log-likelihood approach via EM algorithm, which can also be considered as a selection process among all models in such a way to determine which model best matches the observations. It's intractable to directly maximize the log-likelihood function, leading to complex expressions with no closed-form solutions.

The EM framework starts with some initial parameters. Then, we need to accumulate sufficient statistics and find the posterior distribution of the state $p\left(\mathbf{S} \mid \mathbf{X}, \mathbf{\Theta}^{\text{old}}\right)$ by applying forward-backward algorithm in E step. We utilize this posterior distribution to update parameters Θ via maximizing the complete-data likelihood with respect to each parameter in M step. The function $Q\left(\mathbf{\Theta}, \mathbf{\Theta}^{\text{old}}\right)$ can be defined as:

$$Q\left(\mathbf{\Theta}, \mathbf{\Theta}^{\text{old}}\right) = \sum_{\mathbf{S}} p\left(\mathbf{S} \mid \mathbf{X}, \mathbf{\Theta}^{\text{old}}\right) \ln p(\mathbf{X}, \mathbf{S} \mid \mathbf{\Theta}) \tag{67}$$

We introduce $\gamma(s_{nk})$ to denote the marginal posterior distribution of the nth state s_{nk} and $\xi(s_{n-1,j}, s_{nk})$ to define the joint posterior distribution of two successive states $s_{n-1,j}, s_{nk}$ that

x_{n-1}, x_n are emitted from the jth and kth model state respectively.

$$\begin{aligned}\gamma\left(s_{nk}\right) &= P(s_{nk} \mid \mathbf{X}, \boldsymbol{\Theta}) \\ \xi\left(s_{n-1,j}, s_{nk}\right) &= P(s_{n-1,j}, s_{nk} \mid \mathbf{X}, \boldsymbol{\Theta})\end{aligned} \tag{68}$$

where $\gamma\left(s_{nk}\right)$ denotes the conditional probability $p(s_{nk} \mid \mathbf{X}, \theta)$, where $s_{nk} = 1$ if x_n is emitted from the kth model state, and $s_{nk} = 0$, otherwise.

We can make use of the definition of γ and ξ and substitute Eq. (67) with Eq. (68). We obtain $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$ as:

$$\begin{aligned}Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right) = &\sum_{k=1}^{K} \gamma\left(s_{1k}\right) \ln \pi_k + \sum_{n=2}^{N} \sum_{j=1}^{K} \sum_{k=1}^{K} \xi\left(s_{n-1,j}, s_{nk}\right) \ln A_{jk} \\ &+ \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma\left(s_{nk}\right) \ln p\left(\tilde{\mathbf{x}}_n \mid \Lambda_{nk}\right)\end{aligned} \tag{69}$$

4.3 BAGMM Integration into the HMM framework

From the previous section, the emission distribution $p\left(\tilde{\mathbf{x}}_n \mid \Lambda_{nk}\right)$ is often taken as Gaussian mixture model (GMM) for most continuous observations cases. 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. So, we suggest integrating the bounded asymmetric Gaussian mixture model (BAGMM) into the HMM framework. The primary motivation behind this choice is the bounded range support from BAGMMM and its asymmetric nature for modeling non-symmetric real-world data. The BAGMM is flexible and has good capabilities to model both symmetric and asymmetric data.

By replacing the emission probability distribution as BAGMM, we can integrate BAGMM into the HMM framework, which is to substitute $p\left(\tilde{\mathbf{x}}_n \mid \Lambda_{nk}\right)$ with Eq. (65) in Eq. (69). In the E step, we obtain $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$ using Eq. (69). In the M step, we maximize $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$ with respect to the parameters $\Theta = \{\boldsymbol{\pi}, \mathbf{A}, \Lambda\}$ in which we treat γ, ξ as a constant. The details are discussed in the subsection.

4.3.1 Estimation of π and A

Using Lagrange multipliers, the maximization concerning π_k and A_{jk} gives the following:

$$\pi_k = \frac{\gamma\left(s_{1k}\right)}{\sum_{j=1}^{K}\gamma\left(s_{1j}\right)} \tag{70}$$

$$A_{jk} = \frac{\sum_{n=2}^{N}\xi\left(s_{n-1,j}, s_{nk}\right)}{\sum_{l=1}^{K}\sum_{n=2}^{N}\xi\left(s_{n-1,j}, s_{nl}\right)} \tag{71}$$

Note that the initialization for π_k and A_{jk} should respect the summation constraints, $\sum_{k=1}^{K}\pi_k = 1$ and $\sum_{k=1}^{K}A_{jk} = 1$.

4.3.2 Estimation of Λ

To maximize $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)$ with respect to Λ_k, we note that the final term in Eq. (69) depends on Λ_k. The Λ_k is a set of parameters of the kth state emission probability distribution, $\Lambda_k = [p_1, \ldots, p_m, \mu_1, \ldots, \mu_m, \sigma_{l1}, \ldots, \sigma_{lm}, \sigma_{r1}, \ldots, \sigma_{rm}]$. Here, we denote $\varphi_n(k,m)$ the probability of being at state s_k at time n with respect to the mth bounded asymmetric Gaussian mixture. According to [85, 86], the $\varphi_n(k,m)$ can be computed as:

$$\varphi_n(k,m) = \frac{\alpha\left(s_{nk}\right)\beta\left(s_{nk}\right)}{\sum_{k=1}^{K}\alpha\left(s_{nk}\right)\beta\left(s_{nk}\right)} \cdot \frac{p(\tilde{\mathbf{x}}_n|\xi_{km})p_{km}}{\sum_{m=1}^{M}p(\tilde{\mathbf{x}}_n|\xi_{km})p_{km}} \tag{72}$$

where $\alpha\left(\mathbf{s}_n\right)$ denotes the joint probability of observing all of the given data up to time n and the hidden state s_n, whereas $\beta\left(\mathbf{s}_n\right)$ represents the conditional probability of all future data from time $n+1$ up to N given the hidden state of s_n:

$$\alpha\left(\mathbf{s}_n\right) \equiv p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{s}_n\right) \tag{73}$$

$$\beta\left(\mathbf{s}_n\right) \equiv p\left(\mathbf{x}_{n+1}, \ldots, \mathbf{x}_N \mid \mathbf{s}_n\right) \tag{74}$$

The mixing coefficient p_{km}^{new} of the mth bounded Asymmetric Gaussian mixture in the state k is given by:

$$p_{km}^{new} = \frac{\sum_{n=1}^{N}\varphi_n(k,m)}{\sum_{n=1}^{N}\sum_{m=1}^{M}\varphi_n(k,m)} \tag{75}$$

The mean $\boldsymbol{\mu}_{kmd}^{new}$ can be defined using the same approach.

$$\boldsymbol{\mu}_{kmd}^{new} = \frac{\sum_{n=1}^{N} \varphi_n(k,m) \left\{ \mathbf{x}_{nd} - \frac{\int_{\partial_{km}} f(\mathbf{u}|\xi_{km})(\mathbf{u}-\mu_{kmd})d\mathrm{u}}{\int_{\partial_{km}} f(\mathbf{u}|\xi_{km})d\mathbf{u}} \right\}}{\sum_{n=1}^{N} \varphi_n(k,m)} \tag{76}$$

Note that in Eq. (76), the term $\int_{\partial_{km}} f(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})du$ is the expectation of function $(\mathrm{u}-\mu_{kmd})$ under the probability distribution $f(\mathbf{x}_d|\xi_{km})$. Then, this expectation can be approximated as:

$$\int_{\partial_{km}} f(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})d\mathrm{u} \approx \frac{1}{M} \sum_{m=1}^{M} (m_{kmd}-\mu_{kmd})\mathrm{H}(m_{kmd}|\Omega_{km}) \tag{77}$$

where $m_{kmd} \sim f(\mathrm{u}|\xi_{km})$ is a set of random variables drawn from the asymmetric Gaussian distribution for the particular component m of the mixture model at the state k. The term $\int_{\partial_{km}} f(\mathrm{u}|\xi_{km})d\mathrm{u}$ in Eq. (76) can be approximated as:

$$\int_{\partial_{km}} f(\mathrm{u}|\xi_{km})d\mathrm{u} \approx \frac{1}{M} \sum_{m=1}^{M} \mathrm{H}(m_{kmd}|\Omega_{km}) \tag{78}$$

and

$$\boldsymbol{\mu}_{kmd}^{new} = \frac{\sum_{n=1}^{N} \varphi_n(k,m) \left\{ \mathbf{x}_{nd} - \frac{\sum_{m=1}^{M} (m_{kmd}-\mu_{kmd})\mathrm{H}(m_{kmd}|\Omega_{km})}{\sum_{m=1}^{M} \mathrm{H}(m_{kmd}|\Omega_{km})} \right\}}{\sum_{n=1}^{N} \varphi_n(k,m)} \tag{79}$$

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

$$\sigma_{l_{kmd}}^{new} = \sigma_{l_{kmd}}^{old} - \left[\left(\frac{\partial^2 Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{l_{kmd}}^2} \right)^{-1} \left(\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{l_{kmd}}} \right) \right] \tag{80}$$

where the first derivative of the model's complete data log-likelihood with respect to left standard deviation $\sigma_{l_{kmd}}$ is given as follows:

$$\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{l_{kmd}}} = 0 \tag{81}$$

$$
\begin{aligned}
\frac{\partial Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{l_{kmd}}} &= \frac{\partial}{\partial \sigma_{l_{kmd}}}\sum_{n=1}^{N}\varphi_n(k,m)\times \\
&\left\{\log p_{km} + \log f(\tilde{\mathbf{x}}_n|\xi_{km}) + \log \mathrm{H}(\tilde{\mathbf{x}}_n|km) - \log \int_{\partial_{km}} f(\vec{\mathrm{u}}|\xi_{km})d\mathrm{u}\right\} \\
&= \frac{\partial}{\partial \sigma_{l_{kmd}}}\sum_{n=1}^{N}\varphi_n(k,m)\left\{\log f(\tilde{\mathbf{x}}_n|\xi_{km}) - \log \int_{\partial_{km}} f(\vec{\mathrm{u}}|\xi_{km})d\mathrm{u}\right\} \\
&= \sum_{n=1,\mathbf{x}_{nd}<\mu_{kmd}}^{N}\varphi_n(k,m)\left(\frac{(\mathbf{x}_{nd}-\mu_{kmd})^2}{\sigma^3_{l_{kmd}}}\right) \\
&- \sum_{i=1,\mathbf{x}_{nd}<\mu_{jd}}^{N}\frac{\varphi_n(k,m)}{\sigma^3_{l_{kmd}}}\left\{\frac{\int_{\partial_{km}} \mathrm{g}_1\left(\mathrm{u}\mid\xi_{km}\right)\left(\mathrm{u}-\mu_{kmd}\right)^2 du}{\int_{\partial_{km}} \mathrm{g}_1\left(\mathrm{u}\mid\xi_{km}\right) du}\right\}
\end{aligned}
\tag{82}
$$

The term $\int_{\partial_{km}} \mathrm{g}_1(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})^2 d\mathrm{u}$ can be approximated as below:

$$
\int_{\partial_{km}} \mathrm{g}_1(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})^2 du \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{l}_{kmd}-\mu_{kmd})^2\mathrm{H}(\mathrm{l}_{kmd}|\Omega_{km}) \tag{83}
$$

where $\mathrm{l}_{kmd} \sim \mathrm{g}_1(\tilde{\mathbf{x}}_n|\xi_{km})$ is a set of random variables drawn from the asymmetric Gaussian distribution with $\mathrm{u} < \mu_{kmd}$ for the particular component m of the mixture model at the state k. Similarly, the term $\int_{\partial_{km}} \mathrm{g}_1(\mathrm{u}|\xi_{km})d\mathrm{u}$ in Eq. (82) can be approximated as:

$$
\int_{\partial_{km}} \mathrm{g}_1(\mathrm{u}|\xi_{km})du \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{l}_{kmd}|\Omega_{km}) \tag{84}
$$

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:

$$
\begin{aligned}
&\frac{\partial^2 Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial\sigma_{l_{kmd}}^2} = -3\sum_{n=1,\mathbf{x}_{nd}<\mu_{kmd}}^{N}\varphi_n(k,m)\left(\frac{\left(\mathbf{x}_{nd}-\mu_{kmd}\right)^2}{\sigma_{l_{kmd}}^4}\right)\\
&-\sum_{n=1,\mathbf{x}_{nd}<\mu_{jd}}^{N}\varphi_n(k,m)\left(\frac{-2}{\sigma_{l_{kmd}}^3\left(\sigma_{l_{kmd}}+\sigma_{r_{kmd}}\right)}\right)\times\\
&\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{kmd}-\mu_{kmd}\right)^2\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}\right\}\\
&-\sum_{n=1,\mathbf{x}_{nd}<\mu_{kmd}}^{N}\frac{\varphi_n(k,m)}{\sigma_{l_{kmd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{kmd}-\mu_{kmd}\right)^4\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}\right\}\\
&-\sum_{n=1,\mathbf{x}_{nd}<\mu_{kmd}}^{N}\frac{-3\varphi_n(k,m)}{\sigma_{l_{kmd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{kmd}-\mu_{kmd}\right)^2\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)}\right\}\\
&-\sum_{n=1,\mathbf{x}_{nd}<\mu_{kmd}}^{N}\frac{\varphi_n(k,m)}{\sigma_{l_{kmd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{l}_{kmd}-\mu_{kmd}\right)^2\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{l}_{kmd}\mid\Omega_{km}\right)\right)^2}\right\}
\end{aligned}
\tag{85}
$$

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

$$
\sigma_{r_{kmd}}^{new} = \sigma_{r_{kmd}}^{old} - \left[\left(\frac{\partial^2 Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial\sigma_{r_{kmd}}^2}\right)^{-1}\left(\frac{\partial Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial\sigma_{r_{kmd}}}\right)\right] \tag{86}
$$

Similar approximations are used for $\frac{\partial Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})}{\partial \sigma_{r_{kmd}}}$ as following:

$$\begin{aligned}\frac{\partial Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{r_{kmd}}} &= \frac{\partial}{\partial \sigma_{r_{kmd}}}\sum_{n=1}^{N}\varphi_n(k,m)\times \\ &\left\{\log p_{km} + \log f(\tilde{\mathbf{x}}_n|\xi_{km}) + \log \mathrm{H}(\tilde{\mathbf{x}}_n|km) - \log \int_{\partial_{km}} f(\vec{\mathrm{u}}|\xi_{km})d\mathrm{u}\right\} \\ &= \frac{\partial}{\partial \sigma_{r_{kmd}}}\sum_{n=1}^{N}\varphi_n(k,m)\left\{\log f(\tilde{\mathbf{x}}_n|\xi_{km}) - \log \int_{\partial_{km}} f(\vec{\mathrm{u}}|\xi_{km})d\mathrm{u}\right\} \\ &= \sum_{n=1,\mathbf{x}_{nd}\geq\mu_{kmd}}^{N}\varphi_n(k,m)\left(\frac{(\mathbf{x}_{nd}-\mu_{kmd})^2}{\sigma^3_{r_{kmd}}}\right) \\ &- \sum_{i=1,\mathbf{x}_{nd}\geq\mu_{jd}}^{N}\frac{\varphi_n(k,m)}{\sigma^3_{r_{kmd}}}\left\{\frac{\int_{\partial_{km}}\mathrm{g}_2\left(\mathrm{u}\mid\xi_{km}\right)\left(\mathrm{u}-\mu_{kmd}\right)^2du}{\int_{\partial_{km}}\mathrm{g}_2\left(\mathrm{u}\mid\xi_{km}\right)du}\right\}\end{aligned} \tag{87}$$

The term $\int_{\partial_{km}}\mathrm{g}_2(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})^2d\mathrm{u}$ can be approximated as below:

$$\int_{\partial_{km}}\mathrm{g}_2(\mathrm{u}|\xi_{km})(\mathrm{u}-\mu_{kmd})^2du \approx \frac{1}{M}\sum_{m=1}^{M}(\mathrm{r}_{kmd}-\mu_{kmd})^2\mathrm{H}(\mathrm{r}_{kmd}|\Omega_{km}) \tag{88}$$

where $\mathrm{r}_{kmd} \sim \mathrm{g}_2(\tilde{\mathbf{x}}_n|\xi_{km})$ is a set of random variables drawn from the asymmetric Gaussian distribution with $\mathrm{u} \geq \mu_{kmd}$ for the particular component m of the mixture model at the state k. Similarly, the term $\int_{\partial_{km}}\mathrm{g}_2(\mathrm{u}|\xi_{km})d\mathrm{u}$ in Eq. (87) can be approximated as:

$$\int_{\partial_{km}}\mathrm{g}_2(\mathrm{u}|\xi_{km})d\mathrm{u} \approx \frac{1}{M}\sum_{m=1}^{M}\mathrm{H}(\mathrm{r}_{kmd}|\Omega_{km}) \tag{89}$$

Similar approximations are used for $\frac{\partial^2 Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})}{\partial \sigma_{r_{kmd}}^2}$ as following:

$$\begin{aligned}
&\frac{\partial^2 Q\left(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}\right)}{\partial \sigma_{r_{kmd}}^2} = -3 \sum_{n=1,\mathbf{x}_{nd}\geq \mu_{kmd}}^{N} \varphi_n(k,m)\left(\frac{(\mathbf{x}_{nd}-\mu_{kmd})^2}{\sigma_{r_{kmd}}^4}\right) \\
&- \sum_{n=1,\mathbf{x}_{nd}\geq \mu_{jd}}^{N} \varphi_n(k,m)\left(\frac{-2}{\sigma_{r_{kmd}}^3\left(\sigma_{l_{kmd}}+\sigma_{r_{kmd}}\right)}\right)\times \\
&\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{r}_{kmd}-\mu_{kmd}\right)^2 \mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}\right\} \\
&- \sum_{n=1,\mathbf{x}_{nd}\geq \mu_{kmd}}^{N} \frac{\varphi_n(k,m)}{\sigma_{r_{kmd}}^6}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{r}_{kmd}-\mu_{kmd}\right)^4 \mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}\right\} \\
&- \sum_{n=1,\mathbf{x}_{nd}\geq \mu_{kmd}}^{N} \frac{-3\varphi_n(k,m)}{\sigma_{r_{kmd}}^4}\left\{\frac{\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{r}_{kmd}-\mu_{kmd}\right)^2 \mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}{\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)}\right\} \\
&- \sum_{n=1,\mathbf{x}_{nd}\geq \mu_{kmd}}^{N} \frac{\varphi_n(k,m)}{\sigma_{r_{kmd}}^6}\left\{\frac{\left(\frac{1}{M}\sum_{m=1}^{M}\left(\mathrm{r}_{kmd}-\mu_{kmd}\right)^2 \mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)\right)^2}{\left(\frac{1}{M}\sum_{m=1}^{M}\mathrm{H}\left(\mathrm{r}_{kmd}\mid \Omega_{km}\right)\right)^2}\right\}
\end{aligned} \tag{90}$$

4.3.3 Complete algorithm

The complete learning of BAGMM-HMM is given in Algorithm 3, where $epoch_{max}$ is the maximum number of iterations. The goal of this algorithm is to find the optimal parameters of $\Theta = \{\boldsymbol{\pi}, \mathbf{A}, \Lambda\}$.

The flowchart of this algorithm is shown in Figure 9. First, we initialize π and transition probability $\mathbf{A}$ with the mean probability according to the number of hidden states and number of mixture components and employ K-Means to initialize parameters of BAGMM. Then, we iterate through the E step and M step until convergence where we accumulate sufficient statistics using the forward-backward algorithm in the E step and update the parameters in the M step.

4.4 Experimental Results

In this section, the effectiveness of our model is tested on some real-world applications, including occupancy estimation and human activity recognition (HAR). We compare our approach (BAGMM-HMM) with asymmetric Gaussian mixture model hidden Markov model (AGMM-HMM),

Algorithm 3 Parameters learning for BAGMM-HMM.

Input:Dataset $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$, $epoch_{max}$.
Output: $\{\boldsymbol{\pi}, \mathbf{A}, \Lambda\}$.
{Initialization for $\Theta = [\pi, A, \Lambda]$**}**:
{Expectation Maximization}:
while iterations $\leq epoch_{max}$ **or** relative changes of parameters **not** converged **do**
  {[**E Step**]}:
  for all $[\vec{X}_1, \ldots, \vec{X}_N]$ do
    Compute $\gamma\left(s_{nk}\right)$ and $\xi\left(s_{n-1,j}, s_{nk}\right)$ using forward-backward algorithm.
    Accumulate sufficient statistics according to Eq. (67)
  {[**M step**]}:
  for all $1 \leq j \leq K$ do
    Update π_k, A_{jk} using Eqs. (70 & 71)
    Update p_{km}^{new}, $\boldsymbol{\mu}_{kmd}^{new}$, $\sigma_{l_{kmd}}^{new}$, $\sigma_{r_{kmd}}^{new}$ & $\vec{\sigma}_{r_j}$ using Eqs. (75, 76, 80, & 86).
  end for
end while

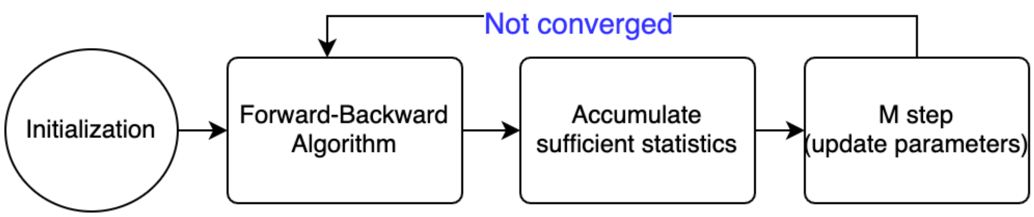

Figure 9: Training process.

bounded Gaussian mixture hidden Markov model (BGMM-HMM), and Gaussian mixture hidden Markov model (GMM-HMM). For comparison, we use the following 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:

$$\left(\frac{TP}{TP + FN}\right)$$

specificity, which is computed as:

$$\left(\frac{TN}{TN + FP}\right)$$

In addition, particularly in case of imbalanced dataset, we must also examine the F1 Score, the harmonic mean of precision and recall, which is computed as:

$$2 \times (precision \times recall)/(precision + recall)$$

G-mean 1, the geometric mean of precision and recall, which is computed as:

$$\sqrt{precision \times recall}$$

G-mean 2, the geometric mean of specificity and recall, which is computed as:

$$\sqrt{specificity \times recall}$$

Mathew' s correlation coefficient (MCC), which is computed as:

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

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

4.4.1 Occupancy Estimation

Indoor occupancy estimation is a critical analytical task for several applications, such as smart buildings or monitoring the energy consumption for power saving. Automating the devices in a building based on occupancy estimation has proved to be very efficient since some research works have indicated that one-third of energy can be saved while using this technique [87, 88].

In terms of privacy, most occupancy detection systems and their modeling approaches avoid employing cameras or audio recorders in favour of non-intrusive sensors, which can be divided into two categories: pyroelectric infrared sensors (PIR) and ambient sensors. For the first category, some research works have been proposed to utilize PIR sensors, and ultrasonic sensors [89, 90]. For the second category, some research works [91, 92] have considered environmental features, such as CO_2 human emission, temperature, humidity, and sound level. Moreover, many machine learning approaches have been used to predict occupants, such as Support Vector Machines (SVM) [90], Logistic Regression [93] and HMMs [94, 95, 96]. They have been utilized to model the extracted features from the environmental data and proved their effectiveness in the occupancy estimation task.

In this section, we employ BAGMM-HMM to estimate occupancy in an office room and hence be the first to tackle this problem with a bounded asymmetric Gaussian mixture-based HMM. Our occupancy estimation task is based on low-cost non-intrusive environmental sensors without bothering privacy policy.

4.4.1.1 Occupancy Detection Dataset

The dataset of the first experiment for occupancy detection is from UCI machine learning Repository [95]. The experimental data about temperature, humidity, light, the ratio of humidity, and CO_2 were obtained from time-stamped pictures taken every minute, which have two labels, occupied and not occupied, respectively. We select training data from two days with 1993 observations and validation data from four days with 4879 observation, for our experiments.

The results in Table 8, showed promising average accuracy for our BAGMM-HMM as compared to AGMM-HMM, BGMM-HMM, and GMM-HMM: 94.90%, 78.30%, 83.58%, and 76.84%, respectively. These results show the effectiveness of our proposed model for occupancy detection. BAGMM-HMM, AGMM-HMM and BGMM-HMM converge faster than traditional GMM-HMM because of bounded range support.

In Figure 10, we present the confusion matrix for this dataset using BAGMM-HMM. Since this is binary classification, our parameters setting is 2 for both the number of hidden states and mixture components. Figure 11 displays the ground truth and our estimated results. From the figures mentioned above, we can see again that our model has an excellent performance.

Table 8: Occupancy detection results using different HMM models.

	HMM Models			
Metrics	BAGMM-HMM	AGMM-HMM	BGMM-HMM	GMM-HMM
Epoch	4	3	3	15
Accuracy	94.90%	78.30%	83.58%	76.84%
Precision	95.83%	88.96%	90.51%	88.59%
Recall	94.90%	78.30%	83.58%	76.84%
Specificity	98.45%	93.70%	91.51%	93.28%
F1-score	95.06%	80.06%	84.80%	78.74%
G-mean 1	95.36%	83.45%	86.97%	82.50%
G-mean 2	96.66%	85.65%	89.22%	84.66%
MCC	87.24%	60.52%	67.49%	58.77%

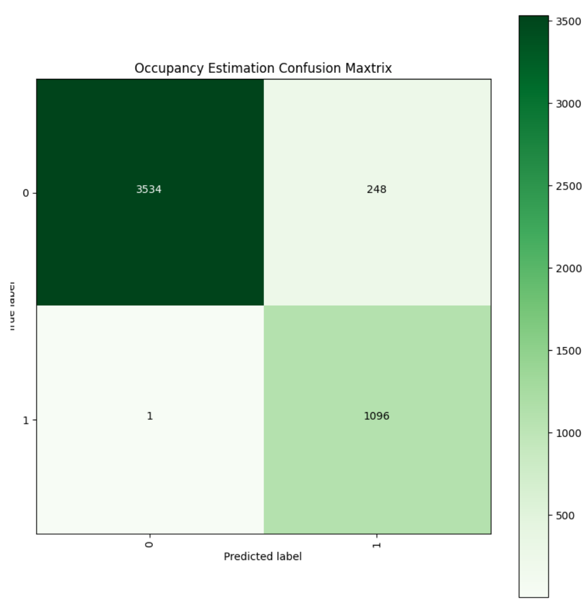

Figure 10: Occupancy detection confusion matrix for BAGMM-HMM.

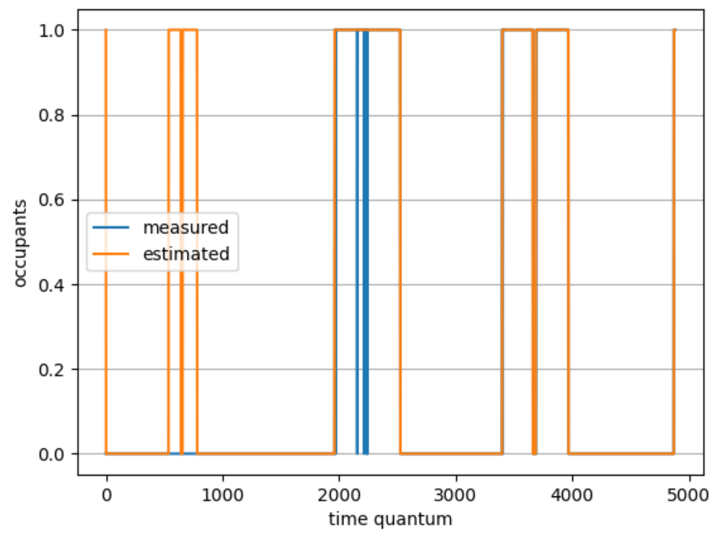

Figure 11: Occupancy detection using BAGMM-HMM.

4.4.1.2 Occupancy estimation dataset

The dataset consists of environmental sensors data collected in an office in Grenoble Institute of Technology, which is housing four people. The dataset comprises luminance, CO_2 concentration, relative humidity (RH), temperature, motion, power consumption, window, door position, and acoustic pressure from a microphone. The data collection is performed continuously with an interval of half an hour. The number of occupants is obtained from recorded videos and used for validation only.

The dataset excludes the timestamp and label of occupants, which is observed information, where the number of occupants is the hidden states that we need to determine. Eight dimensional sensors outputs over a time interval $t = 30$ minutes represent our data and there are 5 hidden states $S = \{s_0, s_1, s_2, s_3, s_4\}$ in this dataset as shown in Figure 12. At time t_0, the number of occupants can be one of the hidden states as shown using green arrows in Figure 12. Each hidden state may switch to another with the transition probability at any time, as shown using black arrows. The red dashed arrows are the emission probabilities indicating the connections between hidden states and observations at a specific time t_n.

With respect to the choice of features, the research paper [97] indicates that the level of CO_2

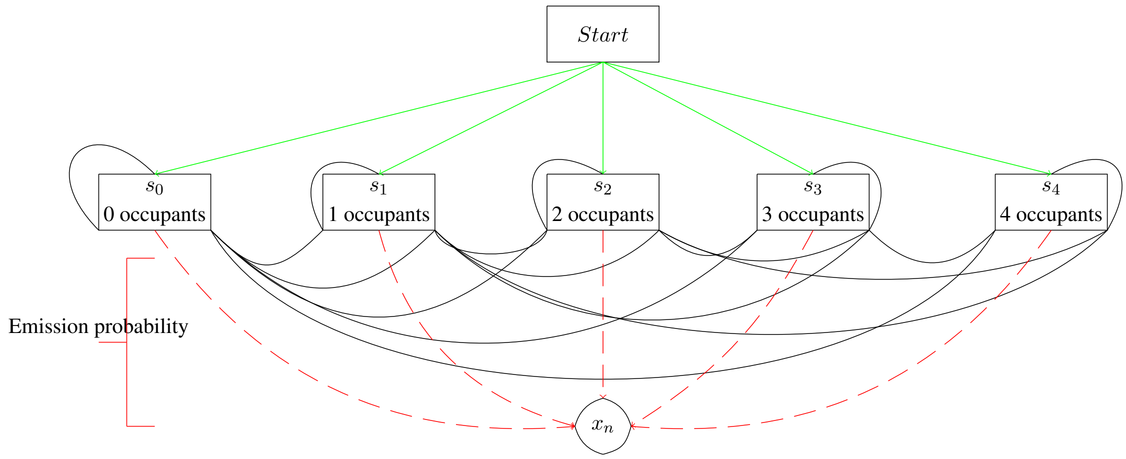

Figure 12: HMM for occupancy estimation according to the case of study.

do not rise immediately as a person comes in, and the authors only employed a subset of features for training: {acoustic pressure, occupancy from power, motion counting}. Another consideration is to re-evaluate the nature of the selected emission probability distribution, which is BAGMM in our work. In this experiment, we use all the features except the datetime and occupancy labels. The BAGMM-HMM is trained according to Algorithm 3 to estimate the model parameters that are employed to test the validation dataset.

4.4.1.3 Experimental Results

The observations in the dataset are collected in the time frame of 20 days every 30 minutes. We choose to train our model using the data collected on days from May 4th, 2015 to May 14th, 2015; test and adjust the model parameters using the data from May 15th, 2015 to May 20th, 2015; validate the model for the rest of data. The compared models are also trained with the same raw data. We just let the models exploit the features and tune the hyperparameters for the models.

After many experiments, the HMM models for our experiments use $K = 5$ for the number of hidden states and $M = 3$ for the number of mixtures to have the best performance. The occupancy estimation comparison results are presented in Table 9. The BAGMM-HMM achieves the best performance with an average accuracy of 86.39% and the highest F1-score with 85.52% compared to 78.45% and 79.28% for AGMM-HMM, 75.42%, and 64.86% for BGMM-HMM against 70.69%, and 75.42% for GMM-HMM, respectively. Our proposed model distinguishes itself as compared to the other models with respect to the considered performance metrics.

Table 9: Occupancy estimation comparison using different HMM models.

	HMM Models			
Metrics	BAGMM-HMM	AGMM-HMM	BGMM-HMM	GMM-HMM
Epoch	4	4	2	10
Accuracy	86.39%	78.45%	75.42%	70.69%
Precision	85.71%	82.91%	56.89%	83.97%
Recall	86.38%	78.45%	75.42%	70.69%
Specificity	75.04%	82.47%	24.57%	88.57%
F1-score	85.52%	79.28%	64.86%	75.42%
G-mean 1	86.05%	80.66%	65.51%	77.05%
G-mean 2	80.52%	80.43%	43.05%	79.13%
MCC	68.35%	57.37%	52.28%	54.39%

The normalized confusion matrix is given in Figure 13. We notice the dataset is an imbalanced dataset from the confusion matrix. But overall, our model can outperform the other HMM models with the same training data.

Figure 14 presents the results obtained from the BAGMM-HMM with 86.39% accuracy, compared with the ground truth as shown with the blue line.

4.4.2 Human Activity Recognition (HAR)

Human activity recognition (HAR) has emerged as an active area of research over the past few years [70, 71] due to many novel ubiquitous applications such as smart buildings, just-in-time surveillance, interactive game interfaces, and home healthcare. The goal of an activity recognition system is to recognize human activities given video clips or environmental sensors data (for privacy concerns) over a time series.

4.4.2.1 HAR Dataset

In this section, we present our experimental results of the proposed model on the challenging human activity recognition (HAR) dataset from UCI machine learning repository [98]. The experiments using this dataset have been carried out with a group of 30 volunteers who performed six activities (walking, walking upstairs, walking downstairs, sitting, standing, laying) wearing a

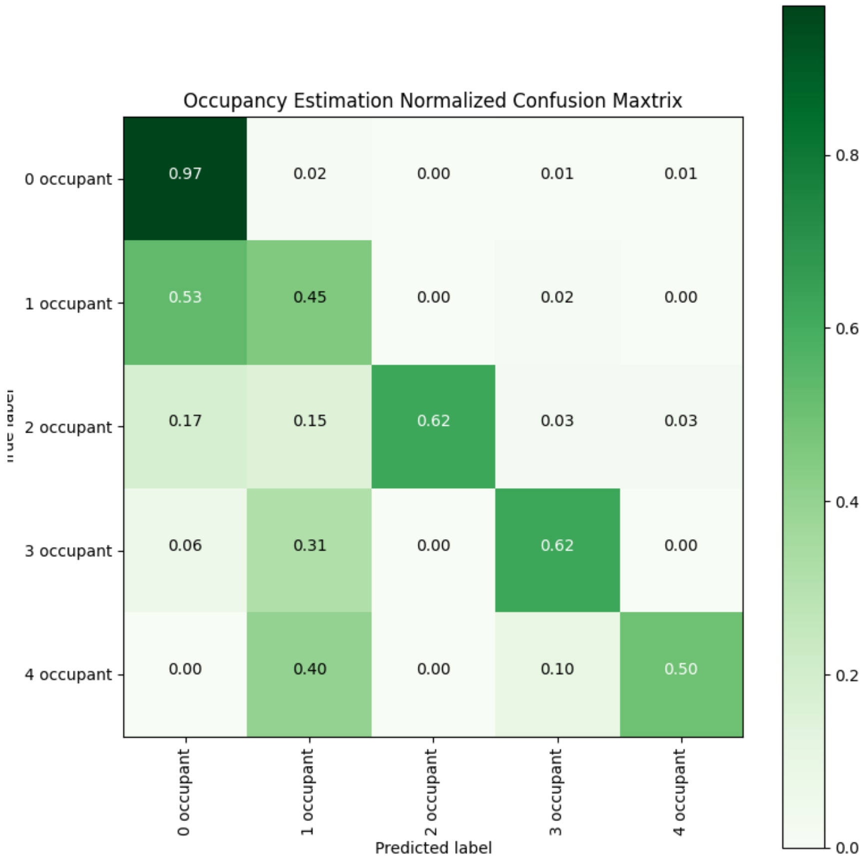

Figure 13: Occupancy estimation normalized confusion matrix for BAGMM-HMM.

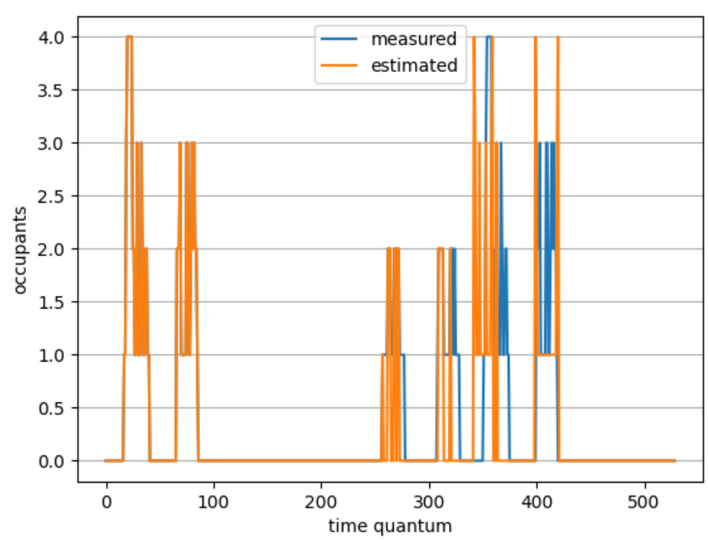

Figure 14: Occupancy estimation using BAGMM-HMM.

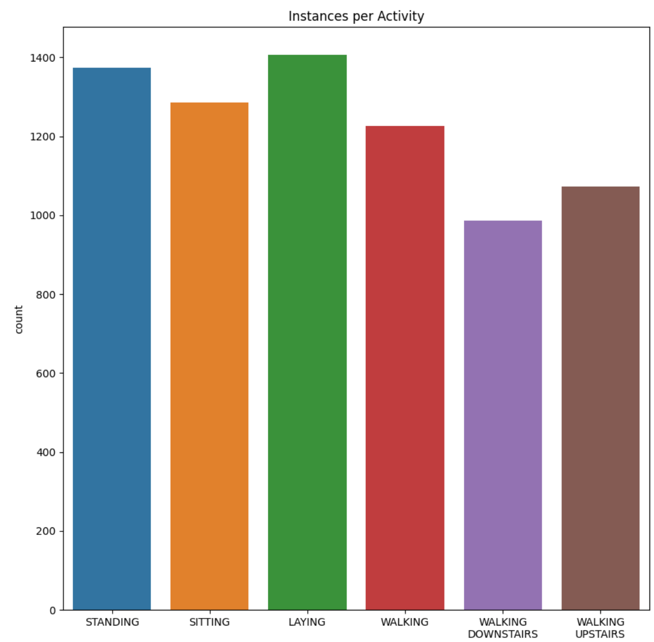

Figure 15: HAR dataset: Instances per activity.

smartphone on the waist. The data comprise 3-axial linear acceleration and 3-axial angular velocity collected by the smartphone's embedded accelerometer and gyroscope at a constant rate of 50Hz. Besides, the experiments have been video-recorded to label the data manually. The dataset was randomly partitioned into two sets, where 70% of the volunteers were selected to generate the training data and 30% for the test data.

4.4.2.2 Preprocessing and Data Visualization

We concatenate all the signal data from the Inertial Signals folder, which has nine files, as our training features. However, the combined features are such a large matrix with a size of 7352×1152 to which we applied principal component analysis (PCA) to reduce the dimension from 1152 to 100. We utilize exploratory data analysis (EDA) to analyze the dataset. We notice that the dataset is balanced, as indicated in Figure 15 that shows the number of data instances per activity.

Furthermore, there are two categorical activities: static (sitting, standing, laying) and dynamic

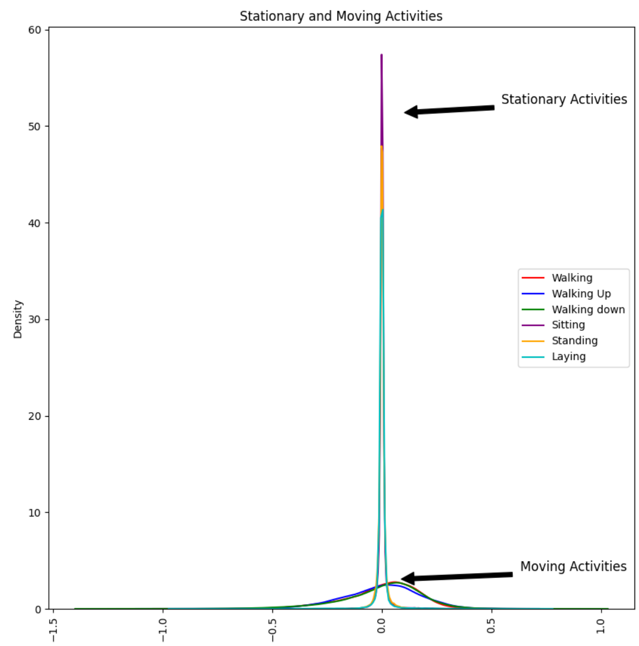

Figure 16: HAR dataset: stationary and moving activities.

(walking, walking upstairs, walking downstairs) activities, respectively. The body acceleration features in the y-axis are significant in stationary activity while not substantial in moving action, as shown in Figure 16.

4.4.2.3 Methodology and Results

An HMM is trained for classifying each human activity using corresponding training data. For the testing stage, the log-likelihood of given testing sensor data is calculated by the respective six trained HMMs, and the class label is assigned according to the maximum likelihood. Our training and predicting process can be observed in Figure 17.

Furthermore, our proposed model outperforms other HMMs, with the best configuration being $K = 2$ states and $M = 2$ mixture components associated with each state shown in Table 10. For the sake of time-saving, we decrease the number of draws from the asymmetric Gaussian distribution during the M step from 4,000 to 1,000. The convergence of BAGMM-HMM is faster than the GMM-HMM model. The results obtained with the BAGMM-HMM are indubitably better than

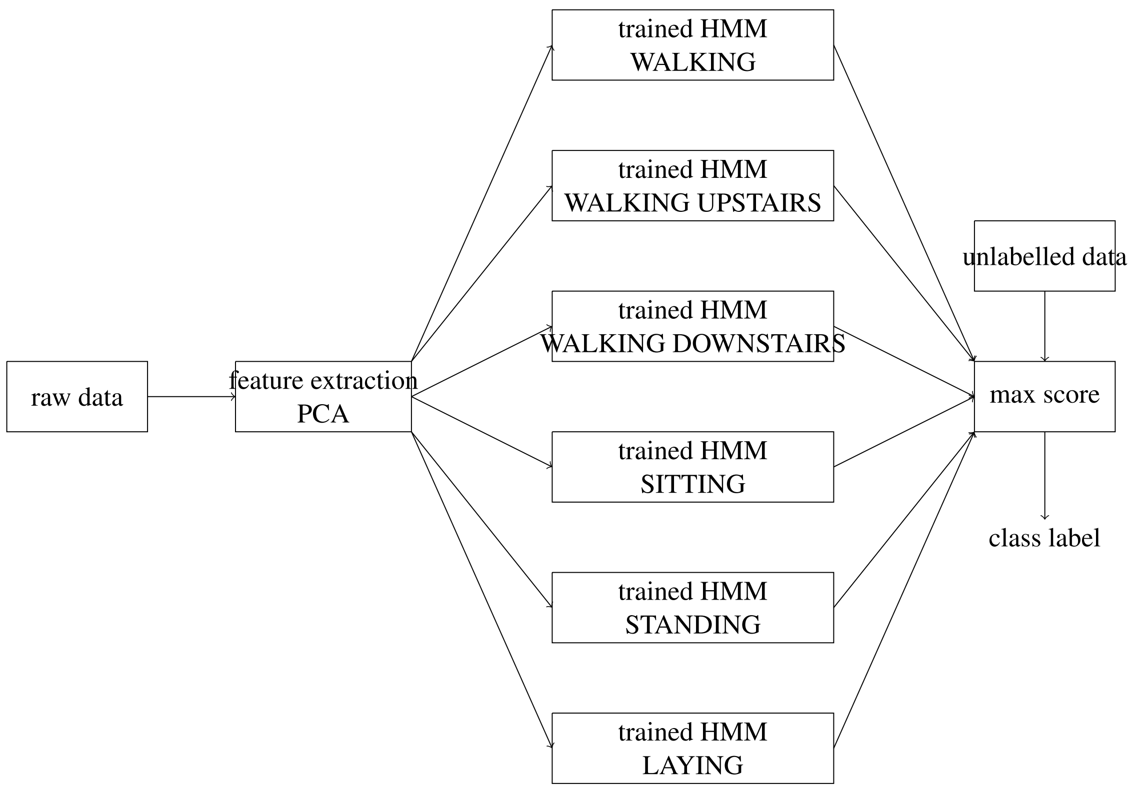

Figure 17: HMM for activity recognition accodring to the case of study.

those with the HMMs, especially the highest accuracy of 84.64% for BAGMM-HMM.

Table 10: Activity recognition results using different HMM models.

	HMM Models			
Metrics	BAGMM-HMM	AGMM-HMM	BGMM-HMM	GMM-HMM
Accuracy	84.62%	77.27%	76.92%	75.00%
Precision	92.31%	69.32%	70.94%	69.44%
Recall	84.62%	77.27%	76.92%	75.00%
Specificity	97.20%	95.24%	24.57%	95.00%
F1-score	83.44%	71.21%	71.64%	68.88%
G-mean 1	88.38%	73.19%	73.87%	72.16%
G-mean 2	90.69%	85.79%	85.45%	84.40%
MCC	83.93%	69.98%	70.16%	68.02%

Chapter 5

Conclusion

This thesis is based on the bounded asymmetric Gaussian mixture model (BAGMM) and its model selection criterion, as well as its integration into the framework of hidden Markov models (HMM). The fact that BAGMM generally performs better than the AGMM is due to its bounded support range which motivated us to further explore the extent of this distribution on various challenging applications.

Chapter 2 discusses 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. The proposed model is applied to synthetic datasets, real datasets and an application is developed for occupancy detection. From all the experimental results, it is observed that the BAGMM and the MML provide strong modeling ability for high-dimensional and complex datasets.

Then chapter 3 proposes 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 dimensionality 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.

Finally, we presented a new extension to the traditional HMM by modifying its emission probability distribution as bounded asymmetric Gaussian mixture. The main goal was to enhance HMM capability of modeling non-symmetric data with bounded support without performing major modifications on its underlying conventional structure. It's examined from all real-life applications that we have performed that the proposed model outperforms all the comparable Gaussian mixture-based HMMs, including the AGMM-HMM, BGMM-HMM, and the traditional Gaussian mixture-based HMM. The particular motivation in adopting bounded asymmetric Gaussian mixtures as the emission probability distribution is encouraged by their sound mathematical foundation and excellent capabilities to approximate and model diverse shapes of real-world data.

Future works could be devoted to applying the proposed frameworks to other challenging applications or considering other learning techniques such as Bayesian inference or variational Bayes. Finally other bounded distributions could also be integrated into the framework of HMMs.

Bibliography

[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.

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

[3] 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.

[4] 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. 013005–013005, 2008.

[5] M. Allili, "Wavelet Modeling Using Finite Mixtures of Generalized Gaussian Distributions: Application to Texture Discrimination and Retrieval," *Image Processing, IEEE Transactions on*, vol. 21, pp. 1452–1464, April 2012.

[6] 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.

[7] 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.

[8] T. Elguebaly and N. Bouguila, "Bayesian learning of generalized gaussian mixture models on biomedical images," in *Artificial Neural Networks in Pattern Recognition, 4th IAPR TC3*

Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings (F. Schwenker and N. E. Gayar, eds.), vol. 5998 of *Lecture Notes in Computer Science*, pp. 207–218, Springer, 2010.

[9] T. Elguebaly and N. Bouguila, "Infinite generalized gaussian mixture modeling and applications," in *Image Analysis and Recognition - 8th International Conference, ICIAR 2011, Burnaby, BC, Canada, June 22-24, 2011. Proceedings, Part I* (M. Kamel and A. C. Campilho, eds.), vol. 6753 of *Lecture Notes in Computer Science*, pp. 201–210, Springer, 2011.

[10] T. Elguebaly and N. Bouguila, "A nonparametric bayesian approach for enhanced pedestrian detection and foreground segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2011, Colorado Springs, CO, USA, 20-25 June, 2011*, pp. 21–26, IEEE Computer Society, 2011.

[11] T. Elguebaly and N. Bouguila, "Generalized gaussian mixture models as a nonparametric bayesian approach for clustering using class-specific visual features," *J. Vis. Commun. Image Represent.*, vol. 23, no. 8, pp. 1199–1212, 2012.

[12] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors," *IEEE transactions on Information Theory*, vol. 45, no. 3, pp. 909–919, 1999.

[13] Y. Bazi, L. Bruzzone, and F. Melgani, "An unsupervised approach based on the generalized gaussian model to automatic change detection in multitemporal sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 4, pp. 874–887, 2005.

[14] S. Kasaei, M. Deriche, and B. Boashash, "A novel fingerprint image compression technique using wavelets packets and pyramid lattice vector quantization," *IEEE Transactions on Image Processing*, vol. 11, no. 12, pp. 1365–1378, 2002.

[15] S. G. Chang, B. Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE transactions on image processing*, vol. 9, no. 9, pp. 1532–1546, 2000.

[16] 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.

[17] P. Hedelin and J. Skoglund, "Vector quantization based on gaussian mixture models," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 385–401, Jul 2000.

[18] J. Lindblom and J. Samuelsson, "Bounded Support Gaussian Mixture Modeling of Speech Spectra," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 88–99, Jan 2003.

[19] M. Azam and N. Bouguila, "Speaker verification using adapted bounded gaussian mixture model," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 300–307, IEEE, 2018.

[20] T. M. Nguyen, Q. Jonathan Wu, and H. Zhang, "Bounded generalized gaussian mixture model," *Pattern Recognition*, vol. 47, no. 9, pp. 3132–3142, 2014.

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

[22] M. Azam and N. Bouguila, "Blind source separation as pre-processing to unsupervised keyword spotting via an ica mixture model," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 833–836, IEEE, 2018.

[23] M. Azam, B. Alghabashi, and N. Bouguila, *Multivariate Bounded Asymmetric Gaussian Mixture Model*, pp. 61–80. Cham: Springer International Publishing, 2020.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[25] 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.

[26] 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.

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

[28] 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.

[29] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, December 1974.

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

[31] 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.

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

[33] 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*, pp. 54–69, Springer, 1999.

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

[35] A. Kołcz, X. Sun, and J. Kalita, “Efficient handling of high-dimensional feature spaces by randomized classifier ensembles,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 307–313, 2002.

[36] 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.

[37] 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.

[38] 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), pp. 364–369, Springer Berlin Heidelberg, 2011.

[39] 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.

[40] 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.

[41] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[42] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology," *Bulletin of the American Mathematical Society*, vol. 73, no. 3, pp. 360–363, 1967.

[43] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.

[44] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes," *Inequalities*, vol. 3, no. 1, pp. 1–8, 1972.

[45] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[46] M. Bicego, U. Castellani, and V. Murino, "A hidden markov model approach for appearance-based 3d object recognition," *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2588–2599, 2005.

[47] H. Lee, D. Lee, and H.-J. Lee, "A predictive initialization of hidden state parameters in a hidden markov model for hand gesture recognition," in *2018 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, pp. 206–212, IEEE, 2018.

[48] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[49] S. A. Frank, "The common patterns of nature," *Journal of evolutionary biology*, vol. 22, no. 8, pp. 1563–1585, 2009.

[50] Z. Xian, M. Azam, M. Amayri, and N. Bouguila, "Model selection criterion for multivariate bounded asymmetric gaussian mixture model," in *EUSIPCO*, 2021.

[51] Z. Xian, M. Azam, and N. Bouguila, "Statistical modeling using bounded asymmetric gaussian mixtures: Application to human action and gender recognition," in *2021 IEEE International Conference on Information Reuse and Integration (IRI)*, IEEE, 2021.

[52] Z. Xian, M. Azam, M. Amayri, W. Fan, and N. Bouguila, "Bounded asymmetric gaussian mixture-based hidden markov models," in *Hidden Markov Models and Applications* (N. Bouguila, W. Fan, and M. Amayri, eds.), Springer, 2021.

[53] T. Kato, S. Omachi, and H. Aso, "Asymmetric gaussian and its application to pattern recognition," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 405–413, Springer, 2002.

[54] T. Elguebaly and N. Bouguila, "Background subtraction using finite mixtures of asymmetric gaussian distributions and shadow detection," *Mach. Vis. Appl.*, vol. 25, no. 5, pp. 1145–1162, 2014.

[55] T. Elguebaly and N. Bouguila, "Simultaneous high-dimensional clustering and feature selection using asymmetric gaussian mixture models," *Image Vis. Comput.*, vol. 34, pp. 27–41, 2015.

[56] S. Fu and N. Bouguila, "Bayesian learning of finite asymmetric gaussian mixtures," in *Recent Trends and Future Technology in Applied Intelligence - 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings* (M. Mouhoub, S. Sadaoui, O. A. Mohamed, and M. Ali, eds.), vol. 10868 of *Lecture Notes in Computer Science*, pp. 355–365, Springer, 2018.

[57] S. Fu and N. Bouguila, "Asymmetric gaussian-based statistical models using markov chain monte carlo techniques for image categorization," in *17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, Orlando, FL, USA, December 17-20, 2018* (M. A. Wani, M. M. Kantardzic, M. S. Mouchaweh, J. Gama, and E. Lughofer, eds.), pp. 1205–1208, IEEE, 2018.

[58] S. Fu and N. Bouguila, "A bayesian intrusion detection framework," in *2018 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2018, Glasgow, Scotland, United Kingdom, June 11-12, 2018*, pp. 1–8, IEEE, 2018.

[59] S. Fu and N. Bouguila, "Asymmetric gaussian mixtures with reversible jump MCMC," in *2018 IEEE Canadian Conference on Electrical & Computer Engineering, CCECE 2018, Quebec, QC, Canada, May 13-16, 2018*, pp. 1–4, IEEE, 2018.

[60] S. Fu and N. Bouguila, "A soft computing model based on asymmetric gaussian mixtures and bayesian inference," *Soft Comput.*, vol. 24, no. 7, pp. 4841–4853, 2020.

[61] 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.

[62] 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, pp. 1133–1142, Nov 1998.

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

[64] 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*, pp. 1–8, June 2007.

[65] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

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

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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*, pp. 789–792, 2014.

[77] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *International Conference on Computer Vision*, sep 2009.

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

[79] 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.

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

[81] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[82] M. Bicego, U. Castellani, and V. Murino, "A hidden markov model approach for appearance-based 3d object recognition," *Pattern Recognition Letters*, vol. 26, no. 16, pp. 2588–2599, 2005.

[83] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[84] E. Andrade, S. Blunsden, and R. Fisher, "Hidden markov models for optical flow analysis in crowds," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, pp. 460–463, 2006.

[85] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[86] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[87] J. Brooks, S. Kumar, S. Goyal, R. Subramany, and P. Barooah, "Energy-efficient control of under-actuated hvac zones in commercial buildings," *Energy and Buildings*, vol. 93, pp. 160–168, 2015.

[88] V. L. Erickson, M. Carreira-Perpiñán, and A. E. Cerpa, "Observe: Occupancy-based system for efficient reduction of hvac energy," in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pp. 258–269, 2011.

[89] P. Liu, S.-K. Nguang, and A. Partridge, "Occupancy inference using pyroelectric infrared sensors through hidden markov models," *IEEE Sensors Journal*, vol. 16, no. 4, pp. 1062–1068, 2016.

[90] J. Petersen, N. Larimer, J. A. Kaye, M. Pavel, and T. L. Hayes, "Svm to detect the presence of visitors in a smart home environment," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5850–5853, 2012.

[91] R. Nasfi, M. Amayri, and N. Bouguila, "A novel approach for modeling positive vectors with inverted dirichlet-based hidden markov models," *Knowledge-Based Systems*, vol. 192, p. 105335, 2020.

[92] H. Rahman and H. Han, "Bayesian estimation of occupancy distribution in a multi-room office building based on co 2 concentrations," *Building Simulation*, vol. 11, no. 3, pp. 575–583, 2018.

[93] M. Snyder, M. Freeman, S. Purucker, and C. Pringle, "Using occupancy modeling and logistic regression to assess the distribution of shrimp species in lowland streams, costa rica: Does regional groundwater create favorable habitat?," *Freshwater Science*, vol. 35, pp. 000–000, 10 2015.

[94] M. Amayri, Q.-D. Ngo, S. Ploix, *et al.*, "Bayesian network and hidden markov model for estimating occupancy from measurements and knowledge," in *2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, vol. 2, pp. 690–695, IEEE, 2017.

[95] 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.

[96] B. Dong, B. Andrews, K. P. Lam, M. Höynck, R. Zhang, Y.-S. Chiou, and D. Benitez, "An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network," *Energy and Buildings*, vol. 42, no. 7, pp. 1038–1046, 2010.

[97] M. Amayri, Q.-D. Ngo, and S. Ploix, "Estimating occupancy from measurement and knowledge with bayesian networks," in *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 508–513, IEEE, 2016.

[98] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, *et al.*, "A public domain dataset for human activity recognition using smartphones.," in *Esann*, vol. 3, p. 3, 2013.