



# Bayesian inference for infinite asymmetric Gaussian mixture with feature selection

Ziyang Song<sup>1</sup> · Samr Ali<sup>1</sup> · Nizar Bouguila<sup>1</sup>

Published online: 2 February 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH, DE part of Springer Nature 2021

## Abstract

Data clustering is a fundamental unsupervised learning approach that impacts several domains such as data mining, computer vision, information retrieval, and pattern recognition. In this work, we develop a statistical framework for data clustering which uses Dirichlet processes and asymmetric Gaussian distributions. The parameters of this framework are learned using Markov Chain Monte Carlo inference approaches. We also integrate a feature selection technique to choose the features that are most informative in order to construct an appropriate model in terms of clustering accuracy. This paper reports results based on experiments that concern dynamic textures clustering as well as scene categorization.

**Keywords** Infinite asymmetric Gaussian mixture model · Feature selection · Gibbs sampling · MCMC · Metropolis-Hastings · Background subtraction

## 1 Introduction

Clustering algorithm is a common unsupervised learning methodology for data analysis and has been widely used for uncovering hidden patterns within data. One extensively considered approach in statistical modeling is mixture models. It is capable of clustering data into homogeneous subgroups where the whole model is represented by a weighted sum of the subpopulations' densities. Due to its flexible representations that provide interpretable results, mixture models are adopted in many applications from different domains.

A well-known assumption in using mixture models for statistical analysis is that considering the per components densities follows the widely used Gaussian assumption (Park et al. 2013). However, the Gaussian distribution is not always an appropriate choice since observations shape may not be strictly symmetric. This is especially the case in natural images where the density distribution may be far from the Gaussian (Hyvärinen and Hoyer 2000; Laptev 2009;

Boutemedjet et al. 2010; Elguebaly and Bouguila 2014). Some evolving systems have been proposed for this problem (Andonovski et al. 2018; Škrjanc et al. 2019). For achieving a better approximation, we investigate the use of asymmetric Gaussian distribution (AGD) which is capable of modeling asymmetric data: AGD has left and right variance parameters which control the shape of different parts to better model the asymmetry of data (Elguebaly and Bouguila 2011; Song et al. 2019).

Parameter estimation is one of the challenges required for the use of mixture models. Various algorithms have been studied to achieve this purpose. The expectation maximization (EM) algorithm is a well-known approach to solve such problem (Bouguila and Ziou 2006). Nevertheless, the EM algorithm is a deterministic approach which is not guaranteed to reach a global optimal solution because of its sensitivity to initialization and overfitting. Instead, Bayesian inference may be used which is extensively studied in mixture modeling (Channoufi et al. 2018; Elguebaly and Bouguila 2014). It provides a strong theoretical framework to design clustering algorithms as well as a formal approach to incorporate prior knowledge about the problem. The authors in Fu and Bouguila (2018) recently studied Bayesian learning of asymmetric Gaussian mixture model. In this work, the authors implement Markov Chain Monte Carlo (MCMC) methods that eradicate the dependency between the mixture parameters and components to address over-fitting problems.

✉ Samr Ali  
al\_samr@encs.concordia.ca

Ziyang Song  
zi\_on@encs.concordia.ca

Nizar Bouguila  
nizar.bouguila@concordia.ca

<sup>1</sup> Concordia University, Montreal, Canada

Several studies and research have been devoted to the automatic selection of the components number which best describes the observations. We introduce the Dirichlet process to address the problem of determination of correct components number since it leads to a realization of a mixture model with an unbounded number of components (Antoniak 1974). This can be considered also as a nonparametric Bayesian approach since it allows the number of components to grow to infinity as required to fit the data (Griffin and Steel 2010). In this paper, we are interested in Bayesian non-parametric approaches for modeling, particularly models based on the Dirichlet process (Bouguila and Ziou 2012). The Dirichlet process allows the number of latent variables to grow as necessary to fit the data, but where individual variables still follow parametric distributions. We address the prevalent problem of choosing the correct mixture components number for mixture models, by introducing the Dirichlet process to extend finite mixture model to an infinite one. Thus, we apply a hierarchical Bayesian learning technique for the proposed infinite asymmetric Gaussian mixture model (IAGM).

Theoretically, the more features used to represent data, the better the clustering algorithm is expected to perform. In practice, however, some features can be noisy, redundant, or uninformative thus can hinder the performance of clustering (Boutemedjet et al. 2009; Bouguila 2009). The presence of many irrelevant features introduces a bias and renders homogeneity measures unreliable (Elguebaly and Bouguila 2015). A viable solution is to remove irrelevant features by identifying the best features to the trained model. The process of reducing the number of collected features to a relevant subset of features is known as feature selection. It can increase the performance of models by eliminating noise in the data, improving model interpretation and decreasing the risk of overfitting. Feature selection methods can be broadly divided into three groups: filters, wrappers, and embedded methods (Adams and Beling 2017).

Filter approaches treat feature selection as preprocessing step where the relevance of each feature is evaluated using the dataset alone. Thus, filters only consider the properties of the features regardless of the model. The authors in Krishnan et al. (1996) propose a trimming feature selection technique specific to mixture models based on the Fisher ratio. However, this method does not iterate through the feature space nor simultaneously estimate model parameters and feature subsets. On the other hand, wrapper approaches evaluate feature relevance with regard to the model. In most cases, a model is built with respect to a subset of features and the model's performance is evaluated based on specified criteria. Wrappers then move through the subset space evaluating feature subsets with regard to the evaluation function. The readers are referred to Galimberti et al. (2018); Marbac and Sedki (2017) for further details about wrapper approaches.

Embedded methods simultaneously select features and construct models. Penalized model-based clustering (Pan and Shen 2007; Bouveyron and Brunet-Saumard 2014) and Bayesian methods (Gustafson et al. 2003; Wang and Zhu 2008) are extensively used in many applications. Feature saliency approaches consider feature selection as parameter estimation problems and recast probability distribution as dependent and independent distributions (Elguebaly and Bouguila 2012; Law et al. 2004). Feature saliency is added as new parameters to the conditional distribution of the mixture model and used to find clusters embedded in feature subspace. Because feature saliency represents the probability of belonging to a mixture-dependent distribution. It can be interpreted as the probability that a feature is relevant. In this paper, we propose a feature saliency measure and integrate it into the Bayesian inference framework. Our approach focuses on detecting cluster structure and discriminating feature relevance simultaneously through Bayesian learning.

To summarize, in this paper, we propose a Bayesian inference approach for infinite asymmetric Gaussian mixture (IAGM) models with a simultaneous feature selection framework. The proposed approach better fits data than the traditionally applied Gaussian mixture models in the case of asymmetric data distribution. Extension to an infinite number of mixture components aims to better estimate the data clusters as required. The simultaneous feature selection approach allows for better approximation due to a better choice of features that represent the data with an enhanced ability for the clustering task; i.e. separating the different classes. A potential drawback of the model is computational complexity which is easily remedied with today's immense available computational resources.

The remainder of this paper is organized as follows. Sect. 2 outlines asymmetric mixture model, sets up Dirichlet process and highlights the feature selection algorithm. Sect. 3 presents the Bayesian inference process and complete algorithm for our model. In Sect. 4, we present the validation on dynamic textures clustering and image categorization tasks and compare it with a number of state-of-the-art methods. Finally, Sect. 5 concludes the paper.

## 2 Infinite asymmetric Gaussian mixture model

In this section, we introduce IAGM with feature saliency algorithm. This paper proposes finite mixture model and then extend it to infinite one. We also introduce the concept of feature saliency and represent our model combined with a feature selection algorithm.

### 2.1 Finite mixture model

Assume we have  $N$  observations dataset  $\chi = (X_1, \dots, X_N)$ , where each of observations  $X_i = (X_{i1}, \dots, X_{iD})$  could be represented as a  $D$ -dimensional random variable and it follows asymmetric Gaussian distribution (AGD). The probability density function for dataset  $\chi$  can be written as:

$$p(\chi | \Theta) = \prod_{i=1}^N \sum_{j=1}^M p_j p(X_i | \xi_j) \tag{1}$$

where  $\Theta = (p_1, \dots, p_M, \xi_1, \dots, \xi_M)$  represents the complete set of parameters fully characterizing the mixture model,  $M$  is the number of components,  $\vec{p} = (p_1, \dots, p_M)$  represents the mixing proportions which must be positive and sum to one, and  $\xi_j$  represents the AGD parameters for mixture component  $j$ .

Given AGD parameters for mixture component  $j$ , the AGD density function is defined as:

$$p(X_i | \xi) \propto \prod_{k=1}^D \frac{1}{(S_{ljk})^{-\frac{1}{2}} + (S_{rjk})^{-\frac{1}{2}}} \times \begin{cases} \exp\left[-\frac{S_{ljk}(X_{ik}-\mu_{jk})^2}{2}\right] & \text{if } X_{ik} < \mu_{jk} \\ \exp\left[-\frac{S_{rjk}(X_{ik}-\mu_{jk})^2}{2}\right] & \text{if } X_{ik} \geq \mu_{jk} \end{cases} \tag{2}$$

where  $\xi_j = (\mu_j, S_{lj}, S_{rj})$  is the set of parameters for AGD with  $\mu_j = (\mu_{j1}, \dots, \mu_{jD})$ ,  $S_{lj} = (S_{lj1}, \dots, S_{ljD})$ , and  $S_{rj} = (S_{rj1}, \dots, S_{rjD})$ .  $\mu_{jk}$ ,  $S_{ljk}$  and  $S_{rjk}$  are the mean, the left precision and the right precision for the  $k$ th-dimensional distribution (Fu and Bouguila 2018) respectively. In this paper, we assume each dimension of observation  $X_i$  is independent; hence, its covariance matrix will be diagonal. This assumption leads to a reduction in the computational power during processing and deployment.

We introduce latent indicator variables  $Z, Z_i$  for each observations  $i$  to indicate which mixture component it belongs to.  $Z_i = (Z_{i1}, \dots, Z_{iM})$  where hidden label  $Z_{ij}$  is set to 1 when the observation  $X_i$  is allocated to component  $j$  otherwise 0. The likelihood function of IAGM is given by:

$$p(\chi | Z, \Theta) = \prod_{i=1}^N p(X_i | \xi_j)^{Z_{ij}} \tag{3}$$

For the mixing weight,  $p_j = p(Z_{i=j})$ ,  $j = 1, \dots, M$  indicates the probability that an observation  $X_i$  is associated with component  $j$ . Hence, the missing allocation variable  $Z$  is given a Multinomial prior as follows:

$$p(Z | \vec{p}) \sim \text{Multi}(\vec{p}) = \prod_{j=1}^M p_j^{n_j} \tag{4}$$

where  $n_j = \sum_{i=1}^N I_{Z_i=j}$  is the number of observations allocated to component  $j$ , and function  $I$  is the indicator function. The mixing proportions are assumed to follow a symmetric Dirichlet prior with concentration parameter  $\frac{\alpha}{M}$  (Rasmussen 1999), so that it is considered that all components sharing an equal prior probability. This then can be denoted as follows:

$$p(\vec{p} | \alpha) \sim \text{Dirichlet}\left(\frac{\alpha}{M}, \dots, \frac{\alpha}{M}\right) = \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{M})^M} \prod_{j=1}^M p_j^{\frac{\alpha}{M}-1} \tag{5}$$

The Dirichlet distribution is a conjugate prior of the Multinomial distribution. Due to the conjugacy of  $Z$  and  $\vec{p}$ , we can achieve better inference by integrating out  $\vec{p}$  to obtain the prior of  $Z$  given hyperparameter  $\alpha$ , and then inferring directly the distribution of the latent variables  $Z$ :

$$p(Z | \alpha) = \int p(Z | \vec{p}) p(\vec{p} | \alpha) d\vec{p} = \frac{\Gamma(\alpha)}{\Gamma(N + \alpha)} \prod_{j=1}^M \frac{\Gamma(\frac{\alpha}{M} + n_j)}{\Gamma(\frac{\alpha}{M})} \tag{6}$$

To use the Gibbs sampling technique, it is required to obtain the conditional prior for a single allocation variable  $Z_i$  given all the others. Keeping all the other indicators fixed, we obtain the following conditional prior:

$$p(Z_i = j | \alpha, Z_{-i}) = \frac{n_{-i,j} + \frac{\alpha}{M}}{N - 1 + \alpha} \tag{7}$$

where the subscript  $-i$  indicates all indexes except  $i$ , and  $Z_{-i} = (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_N)$ .  $n_{-i,j}$  is the number of observations excluding  $X_i$  that are allocated to component  $j$ .

### 2.2 Infinite mixture model

We continue to extend the finite mixture model proposed in last section to an infinite mixture model by letting component number  $M \rightarrow \infty$  and updating the posteriors of indicators in Eq. (7). This is achieved by introducing the Dirichlet process to extend to the infinite mixture model (Blei and Jordan 2006):

$$p(Z_i = j | \alpha, Z_{-i}) = \begin{cases} \frac{n_{-i,j}}{N-1+\alpha} & \text{if } n_{-i,j} > 0 \\ \frac{\alpha}{N-1+\alpha} & \text{if } n_{-i,j} = 0 \end{cases} \tag{8}$$

where  $n_{-i,j} > 0$  indicates that component  $j$  is represented. Thus, an observation  $X_i$  is allocated to an existing component with certain probability proportional to the number of observations already associated with this component, while a new component is only proportional to concentration parameter  $\alpha$  and observations number  $N$ . Given the priors, the conditional posteriors are obtained by combining priors with the likelihood:

$$p(Z_i = j | \dots) = \begin{cases} \frac{n_{-i,j}}{N-1+\alpha} \prod_{k=1}^D p(X_{ik} | \xi_{jk}) & \text{if } n_{-i,j} > 0 \\ \frac{\alpha}{N-1+\alpha} \int p(\xi_j | \dots) p(X_i | \xi_j) d\xi_j & \text{if } n_{-i,j} = 0 \end{cases} \tag{9}$$

where the conditional posteriors of unrepresented component is obtained by integrating over hyperparameters and the integral is not analytically intractable. For inferring intractable posteriors, we adopt Algorithm 2 of Neal’s (Neal 2000) which proposes a sampling method to approximate the desired distribution.

Concerning the concentration parameter  $\alpha$ , we consider  $\alpha$  an inverse Gamma prior with parameter  $\kappa$  and  $\eta$ :

$$p(\alpha^{-1} | \kappa, \eta) \sim \Gamma(\kappa, \eta) \tag{10}$$

Given the likelihood of  $\alpha$  in Eq.(6), we obtain the conditional posterior for  $\alpha$  depending on the observations number  $N$  and the components number  $M$

$$p(\alpha | \kappa, \eta) \sim p(Z | \alpha) p(\alpha | \kappa, \eta) \tag{11}$$

### 2.3 Feature saliency

In this section, we introduce the concept of feature saliency and consider the feature selection problem as a parameter estimation problem (Law et al. 2004). It is natural to consider that different features may have different weights for each of the mixture components. Thus, we define feature saliency as the weight of feature importance.

We assume that a feature is relevant if it follows a mixture-dependent distribution AGD. Otherwise, it may be modeled as a mixture-independent background distribution. In our work, we propose a Gaussian assumption for the background distribution. By treating latent relevant indicator  $\phi_i = (\phi_{i1}, \dots, \phi_{iM})$  with  $\phi_{ij} = (\phi_{ij1}, \dots, \phi_{ijD})$ . We could then represent if a given feature is relevant or not. The binary indicator  $\phi_{ijk} = 1$  if feature  $k$  in observations  $X_i$  is relevant for component  $j$ , otherwise  $\phi_{ijk} = 0$ . Thus, we rewrite the probability density function as follows:

$$p(X | \Theta, \xi^{irr}, \Phi) = \prod_{i=1}^N \sum_{j=1}^M p_j \prod_{k=1}^D [p(X_{ik} | \xi_{jk})^{\phi_k} p(X_{ik} | \xi_{jk}^{irr})^{1-\phi_k}] \tag{12}$$

where the  $\xi^{irr} = (\xi_1^{irr}, \dots, \xi_M^{irr})$  represents the set of parameters for background Gaussian distribution with  $\xi_j^{irr} = (\mu_j^{irr}, (S_j^{irr})^{-1})$ ,  $\mu_j = (\mu_{j1}, \dots, \mu_{jD})$ ,  $S_j = (S_{j1}, \dots, S_{jD})$ .  $\mu_{jk}^{irr}$  and  $S_{jk}^{irr}$  represent the mean and precision for the  $k$  dimensional Gaussian distribution, respectively.

Feature saliency defined as  $\vec{\rho} = (\rho_1, \dots, \rho_M)$  such that  $\rho_j = (\rho_{j1}, \dots, \rho_{jD})$ .  $\rho_{jk} = p(\phi_j = 1)$  represents the prior probability that the feature  $k$  is relevant in mixture component  $j$ . Thus, we could recast the likelihood function after introducing the feature saliency  $\vec{\rho}$ . This can be denoted by:

$$p(X_i | \Theta_F) = \sum_{j=1}^M p_j \prod_{k=1}^D (\rho_{jk} p(X_{ik} | \xi_{jk}) + (1 - \rho_{jk}) p(X_{ik} | \xi_{jk}^{irr})) \tag{13}$$

where  $\Theta_F = (\Theta, \vec{\rho}, \xi^{irr})$  is the full set of parameters of the mixture model after introducing feature saliency. Eq. (13) offers sound generative interpretation for our model. First, the model selects the component  $j$  by sampling from a Multinomial distribution with mixing proportion  $(p_1, \dots, p_k)$ . Then, for each feature dimension  $k = 1, \dots, D$ , we follow a Bernoulli distribution with feature saliency  $\rho_{jk}$ ; if successful, we use the relevant mixture component  $p(X_{ik} | \xi_{jk})$  to generate feature  $k$ ; otherwise, the background component  $p(X_{ik} | \xi_{jk}^{irr})$  will be used. Therefore, we could view the model of previous section as special case when all of the features are relevant.

The conditional posteriors of Dirichlet process mixture could be rewritten after bringing feature saliency into model as:

$$p(Z_i = j | \dots) = \begin{cases} \frac{n_{-i,j}}{N-1+\alpha} \prod_{k=1}^D (\rho_{jk} p(X_{ik} | \xi_{jk}) + (1 - \rho_{jk}) p(X_{ik} | \xi_{jk}^{irr})) & \text{if } n_{-i,j} > 0 \\ \frac{\alpha}{N-1+\alpha} \int p(\xi_j | \dots) p(\xi_j^{irr} | \dots) \times p(X_i | \xi_j) d\xi_j & \text{if } n_{-i,j} = 0 \end{cases} \tag{14}$$

We could use these posteriors to generate new components or allocated observations. For latent allocation variable  $Z = (Z_1, \dots, Z_N)$ ,  $p_j = p(Z_i = j)$  represents the prior probability that observation  $X_i$  is associated with component  $j$ . We could obtain the posterior probability that the observation  $X_i$  is allocated to component  $j$  conditional on having observation  $X_i$  to be:

$$\begin{aligned}
 p(Z_i = j | X_i) &= \frac{p(X_i | \Theta_F, Z_i = j)}{p(X_i | \Theta_F)} \\
 &\propto p_j \prod_{k=1}^D (\rho_{jk} p(X_{ik} | \theta_{jk}) + (1 - \rho_{jk}) p(X_{ik} | \theta_{jk}^{irr}))
 \end{aligned}
 \tag{15}$$

Latent relevancy variable  $\phi_{ijk}$  indicates whether the feature  $k$  is relevant for component  $j$  given the observation  $X_i$ .  $\rho_j = p(\phi_{ijk} = 1)$  represents the prior probability that the feature  $k$  is relevant for component  $j$  given observation  $X_i$ . The posterior probability that the feature  $k$  is relevant for component  $j$  conditioned on  $X_i$  is given by:

$$\begin{aligned}
 p(\phi_{ijk} = 1, Z_i = j | X_i) &= p(Z_i = j | X_i) \cdot \\
 &\frac{\rho_{jk} p(X_{ik} | \xi_{jk})}{\rho_{jk} p(X_{ik} | \xi_{jk}) + (1 - \rho_{jk}) p(X_{ik} | \xi_{jk}^{irr})}
 \end{aligned}
 \tag{16}$$

Posteriors for irrelevant features could be deduced in the same way.

$$\begin{aligned}
 p(\phi_{ijk} = 0, Z_i = j | X_i) &= p(Z_i = j | X_i) \cdot \\
 &\frac{(1 - \rho_{jk}) p(X_{ik} | \xi_{jk}^{irr})}{\rho_{jk} p(X_{ik} | \xi_{jk}) + (1 - \rho_{jk}) p(X_{ik} | \xi_{jk}^{irr})}
 \end{aligned}
 \tag{17}$$

The likelihood function of  $\chi$  conditioned on the complete set of mixture parameters can be obtained. It will be used for further Bayesian inference derivation:

$$\begin{aligned}
 p(\chi | Z, \Phi, \xi, \xi^{irr}) &= \prod_{i=1}^N \prod_{k=1}^D [p(X_{ik} | \xi_{jk})^{\phi_k} \cdot \\
 &\quad (X_{ik} | \xi_{jk}^{irr})^{1-\phi_k}]
 \end{aligned}
 \tag{18}$$

### 3 Non-parametric Bayesian inference

In the Bayesian context, the most important step is the determination of the posteriors for inference. In this section, we describe a MCMC-based inference approach to learn the proposed model (recall that MCMC refers to Markov Chain Monte Carlo methods). The goal of inference is to approximate the posteriors of parameters which absorb the information data to update the priors. Thus, we define a hierarchical Bayesian model and use conjugacy to develop the appropriate posteriors. The parameters are inferred based on a MCMC method. The graphical representation is shown in Fig. 1.

#### 3.1 Parameter estimation for $\mu_{jk}$ and $\mu_{jk}^{irr}$

We consider that the relevant and irrelevant mean parameters  $\mu_{jk}$  and  $\mu_{jk}^{irr}$  follow Gaussian priors with common hyperparameters mean  $\lambda$  and precision  $r$  respectively as follows:

$$\begin{aligned}
 p(\mu_{jk} | \lambda, r) &\sim \mathcal{N}(\lambda, r^{-1}) \\
 p(\mu_{jk}^{irr} | \lambda^{irr}, r^{irr}) &\sim \mathcal{N}(\lambda^{irr}, (r^{irr})^{-1})
 \end{aligned}
 \tag{19}$$

where the hyperparameters mean  $\lambda$  and precision  $r$  are considered as common to all components in a specific dimension  $k$ .  $\lambda$  and  $r$  are given Gamma and inverse Gamma prior with the following shape and mean hyperparameters:

$$p(\lambda) \sim \mathcal{N}(e, f) \quad p(r) \sim \gamma(g, h)
 \tag{20}$$

where  $\lambda, \lambda^{irr}, r, r^{irr}$  have same prior forms and we will omit replicated representation for saving space. The conditional posteriors for  $\mu_{jk}$  and  $\mu_{jk}^{irr}$  are obtained by combining the likelihood in Eq. (18) and the priors in Eq. (19).

$$\begin{aligned}
 p(\mu_{jk} | \dots) &\propto p(\mu_{jk} | \lambda, r) p(\chi | Z, \Phi, \xi, \xi^{irr}) \\
 p(\mu_{jk}^{irr} | \dots) &\propto p(\mu_{jk}^{irr} | \lambda^{irr}, r^{irr}) p(\chi | Z, \Phi, \xi, \xi^{irr})
 \end{aligned}
 \tag{21}$$

For the posteriors of hyperparameters  $\lambda$  and  $r$ , Eq. (19) plays the role of likelihood and combined with priors Eq. (20) to obtain:

$$\begin{aligned}
 p(\lambda | \dots) &\propto p(\lambda) \prod_{j=1}^M p(\mu_{jk} | \lambda, r) \\
 p(r | \dots) &\propto p(r) \prod_{j=1}^M p(\mu_{jk} | \lambda, r)
 \end{aligned}
 \tag{22}$$

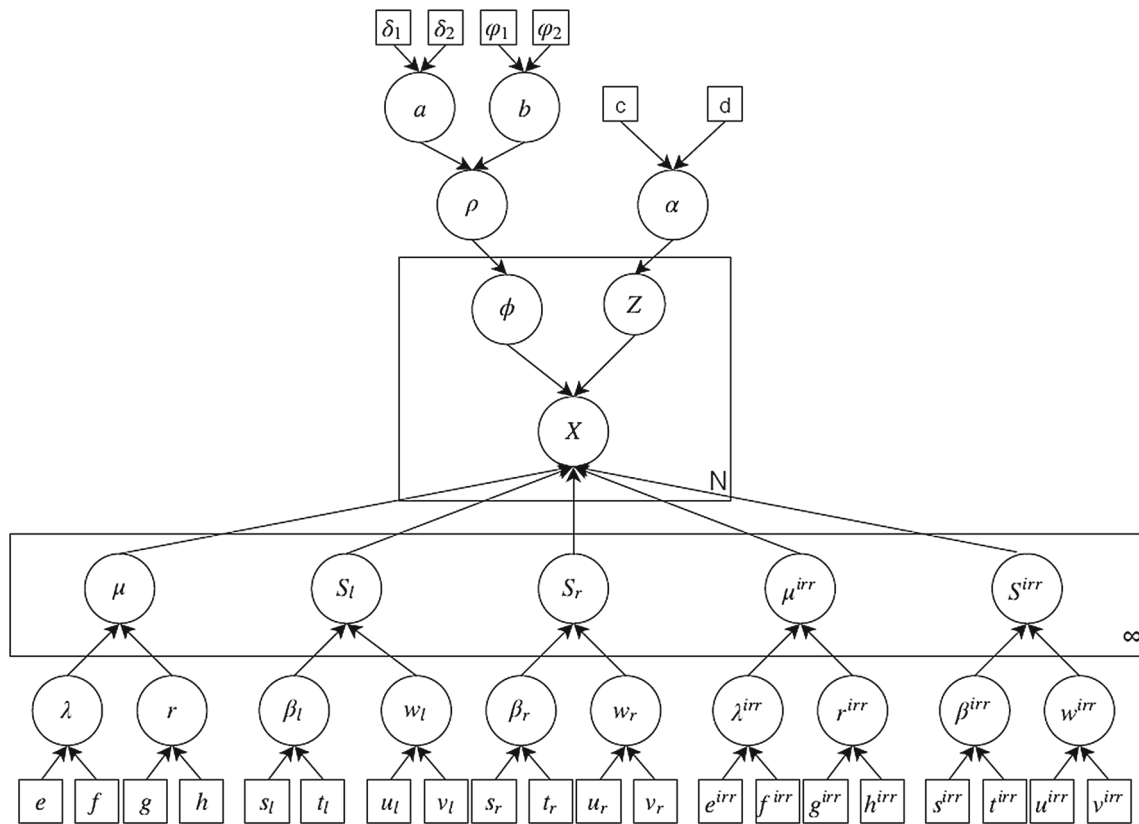
#### 3.2 Parameter estimation for $S_{ljk}, S_{rjk}$ and $S_{jk}^{irr}$

The precision parameters  $S_{ljk}, S_{rjk}$  and  $S_{jk}^{irr}$  are endowed with Gamma priors of common hyperparameters  $\beta$  and  $w$  respectively:

$$\begin{aligned}
 p(S_{ljk} | \beta_l, w_l) &\sim \gamma(\beta_l, w_l^{-1}) \\
 p(S_{rjk} | \beta_r, w_r) &\sim \gamma(\beta_r, w_r^{-1}) \\
 p(S_{jk}^{irr} | \beta^{irr}, w^{irr}) &\sim \gamma(\beta^{irr}, (w^{irr})^{-1})
 \end{aligned}
 \tag{23}$$

where the hyperparameters  $\beta, w$  are common to all components in specific dimension  $k$ .  $\beta$  and  $w$  are given Gamma and





**Fig. 1** Graphical model representation of IAGM. Symbols in circles denote random variables; while the ones in squares denote model parameters. Plates indicate repetition (with the number of repetitions in the lower right), and arcs describe the conditional dependencies between the variables

inverse Gamma priors with the respective shape and mean hyperparameters:

$$p(\beta^{-1}) \sim \gamma(s, t) \quad p(w) \sim \gamma(u, v) \tag{24}$$

where  $\beta_l, \beta_r, \beta^{irr}, w_l, w_r, w^{irr}$  have the same prior forms. The conditional posteriors for  $S_{ljk}, S_{rjk}$  and  $S_{jkk}^{irr}$  are obtained by combining the likelihood in Eq. (18) and the priors in Eq. (23) as follows:

$$\begin{aligned} p(S_{ljk} | \dots) &\propto p(S_{ljk} | \beta_l, w_l) p(\chi | Z, \Phi, \xi, \xi^{irr}) \\ p(S_{rjk} | \dots) &\propto p(S_{rjk} | \beta_r, w_r) p(\chi | Z, \Phi, \xi, \xi^{irr}) \\ p(S_{jkk}^{irr} | \dots) &\propto p(S_j^{irr} | \beta^{irr}, w^{irr}) p(\chi | Z, \Phi, \xi, \xi^{irr}) \end{aligned} \tag{25}$$

For the posteriors of hyperparameters  $\beta$  and  $w$ , Eq. (23) plays the role of likelihood and combined with priors Eq. (24), we can then obtain the following:

$$p(\beta | \dots) \propto p(\beta) \prod_{j=1}^M p(S_{jk} | \beta, w)$$

$$p(r | \dots) \propto p(w) \prod_{j=1}^M p(S_{jk} | \beta, w) \tag{26}$$

### 3.3 Parameter estimation for $\rho$

Feature saliency  $\rho_{jk}$  has support over  $[0, 1]$  and considered naturally as Beta distribution with common hyperparameters  $a$  and  $b$  as following:

$$p(\rho_{jk} | a, b) \sim \text{Beta}(a, b) \tag{27}$$

where the shape hyperparameters  $a$  and  $b$  are common to all components and follows Gamma priors respectively:

$$p(a) \sim \gamma(\delta_1, \delta_2) \quad p(b) \sim \gamma(\varphi_1, \varphi_2) \tag{28}$$

We assume that the latent relevancy parameter  $\phi_{jk}$  are following Bernoulli distribution with probability  $\rho_{jk}$ , so we have:

$$p(\phi_{jk} | \rho_{jk}) \sim \prod_{i=1}^N \rho_{jk}^{\phi_{ijk}} (1 - \rho_{jk})^{(1 - \phi_{ijk})} = \rho_{jk}^{n_{jk}} (1 - \rho_{jk})^{N - n_{jk}} \tag{29}$$

where  $n_{jk} = \sum_{i=1}^N I_{\phi_{jk}=1}$  represents the amount of feature  $k$  relevant for component  $j$  given all of the observations. Considering Eq. (37) as the likelihood, we can obtain the conditional posteriors by multiplying the prior in Eq. (35):

$$p(\rho_{jk} | \dots) \sim p(\phi_{jk} | \rho_{jk})p(\rho_{jk} | a, b) \tag{30}$$

Conditional posteriors can then be obtained by combing Eq. (27) and Eq. (28) as follows:

$$p(a | \dots) \propto p(a) \prod_{j=1}^M p(\rho_{jk} | a, b)$$

$$p(b | \dots) \propto p(b) \prod_{j=1}^M p(\rho_{jk} | a, b) \tag{31}$$

### 3.4 Complete algorithm

Following the inference approach above, we propose a MCMC based algorithm for inferring our hierarchical Bayesian mixture model. Among Monte Carlo methods, Gibbs sampling is one of the most popular methods, and it also widely used for complicated posteriors sampling. We also use Metropolis-Hastings algorithm to generate non-standard posteriors. The Gibbs sequence converges to the joint posterior distribution. The algorithm can be summarized in Algorithm. 1.

## 4 Experimental results

In this section, we validate our algorithm on several challenging experiments; particularly, dynamic textures clustering and scene categorization. We compare our results with multiple state-of-the-art methods of these applications. The hyperparameters chosen are  $e = \mu_y$ ,  $f = \sigma^2$ ,  $g=2$ ,  $h=\frac{2}{\sigma^2}$ ,  $s=0.5$ ,  $t=2$ ,  $u=0.5$ ,  $v=\frac{2}{\sigma^2}$ ,  $\delta_1=2$ ,  $\delta_2=0.5$ ,  $\varphi_1=2$ ,  $\varphi_2=0.5$ ,  $\kappa=0.5$ , and  $\eta=2$ .  $\mu_x$  and  $\sigma_x^2$  are the mean and variance of observations.

### 4.1 Dynamic textures clustering

Dynamic textures are the temporal extension of spatial textures which are defined as sequences of images of moving scenes that exhibit certain stationarity properties in time (seawaves, smoke, foliage, whirlwind) (Doretto et al. 2003). Dynamic textures have drawn tremendous attention during the past years due to their application in several domains in image processing and pattern recognition, such as motion classification, video registration, and computer games (Fan and Bouguila 2013, 2015). In our experiment, we apply the



Fig. 2 Sample frames from the DynTex database

proposed IAGM with simultaneous feature selection for clustering dynamic textures with a representation of LBP-TOP features.

We carry out our experimentation on the challenging dynamic textures dataset; DynTex (Péteri et al. 2010), for evaluating the performance of the algorithm. This dataset contains over 650 dynamic texture video sequences from several categories. In our case, we use a subset of video sequences from 8 different categories: candle, flag, flower, fountain, grass, sea, smoke and tree. Each category has about 20 video sequences. The sample frames from each category are shown in Fig. 2. As a preprocessing step, we extract LBP-TOP descriptors from the selected video sequence. In our experiment, we adopt the parameter choice of 4,4,4,1,1,1 as suggested in Zhao and Pietikainen (2007). The chosen setting of the LBP-TOP descriptor achieves a good performance while it also provides a comparative shorter 48-length feature vector.

Obtained features are modeled using proposed IAGM algorithm. In order to evaluate the performance of the proposed method, we compared our approach with four other methods; infinite Beta-Liouville mixture, infinite generalized Dirichlet mixture, infinite Dirichlet mixture, and infinite Gaussian mixture models. We run these approaches 30 times and get average results for validating the performance. The averages of the clustering accuracy can be observed in Table 1. Fig. 3 shows the confusion matrix for the dataset using IAGM with feature selection. According to the results, IAGM with feature selection approach outperforms the other four methods in terms of the highest categorization accuracy rate (87.02%). It shows significant improvement compared with other methods because it could successfully distinguished 6 categories leading to a higher overall accuracy. The results of dynamic texture clustering demonstrates the advantage of applying mixture models which includes asymmetry characteristics of observations for modelling non-standard shape observations. Meanwhile, simultaneously performing feature selection allows for the inclusion of background noise while accurately representing important features for contributing better performance.

**Algorithm 1** IAGM with feature selection

- 1: **procedure**
- 2: **Initialization:**
- 3: Initialize the truncation level  $K$  and  $T$
- 4: **repeat:**
- 5: Update the latent relevancy variable  $\phi$  from Multivariate Bernoulli distribution with probability  $p(\phi_{ijk} = 1, Z_i = j | X_i)$  in Eq. (16).
- 6: Update mixture parameters  $\mu, \mu^{irr}, S_l, S_r, S^{irr}$  and  $\vec{\rho}$  from conditional posteriors in Eq. (21), Eq. (25) and Eq. (30).
- 7: Update hyperparameters  $\lambda, r, \beta, w, a, b$  from conditional posteriors and update Dirichlet process concentration parameter  $\alpha$  from conditional posteriors in Eq. (22), Eq. (26) and Eq. (31) and Eq. (11).
- 8: Update the latent allocation variable  $Z$  in Eq. (14).
- 9: Update the component number  $M$ .
- 10: The convergence criteria is reached when the difference of the current value of joint posteriors and the previous value is less than  $10^{-4}$ . Otherwise, repeat step 1-5 until convergence.
- 11: **until convergence**

**Table 1** Average accuracy of different algorithms for dynamic textures clustering

Approach	IGM	IDM	IGDM	IBLM	IAGM
Acc (%)	74.87	77.75	80.62	83.37	88.79

grass	1	0	0	0	0	0	0	0	0
sea	0	0.68	0	0	0	0	0	0	0.32
trees	0	0	1	0	0	0	0	0	0
flags	0	0	0	0.53	0	0	0	0.09	0.38
flowers	0	0	0	0	0.96	0	0	0.04	0
foliage	0	0	0	0	0	1	0	0	0
fountains	0	0	0	0	0	0	1	0	0
water	0	0	0	0.03	0	0	0	0.97	0
others	0	0	0	0	0	0	0	0	0

**Fig. 3** Confusion matrix of the IAGM with feature selection for the DynTex database**4.2 Scene categorization**

Humans are proficient at perceiving, recognizing and understanding natural scenes. It have attracted more and more interests to develop machines to simulate human vision functions. The representation of scene images has drawn considerable interests in recent years. In this section, we apply our proposed algorithm to the challenging scene categorization task. Thus, we divide our approach into three parts: feature extraction, image representation, and scene classification.

In this application, we use the UIUC sports event dataset (Li and Fei-Fei 2007) to validate the performance of our algo-

badminton	1	0	0	0	0	0	0	0	0
bocce	0	0.85	0.05	0	0	0	0	0.1	0
croquet	0	0	1	0	0	0	0	0	0
polo	0	0	0.5	0.78	0	0	0.17	0	0
climbing	0	0	0.33	0	0.59	0	0	0	0.08
rowing	0	0	0	0	0	0.57	0.08	0	0.35
sailing	0	0	0	0	0	0.3	0.52	0	0.18
boarding	0.03	0	0	0.05	0.25	0	0	0.67	0
others	0	0	0	0	0	0	0	0	0

**Fig. 4** Confusion matrix of the the IAGM with feature selection for the UIUC sport event dataset

gorithm. This dataset consists of 8 different sport event classes: rowing (250 images), badminton (200 images), polo (182 images), bocce (137 images), snowboarding (190 images), croquet (236 images), sailing (190 images), and rock climbing (194 images). Fig. 5 demonstrates its diverse nature.

We represent each image by a collection of local image patches. Particularly, we adopt scale-invariant feature transform (SIFT) descriptors of  $16 \times 16$  pixel patches computed over a grid with spacing of 8 pixels. Then, we employ bag of visual words (BoVW) approach to have an overall representation of each image. We then use k-means algorithm to cluster our training dataset in a vocabulary of  $V$  visual words. For each SIFT keypoint, it will be allocated to the nearest vocabulary in codebook. The points in the image that can be approximated by each of the visual words. Thus, each image can be represented as a frequency histogram over the  $V$  visual words. Then, we use IAGM with feature selection model to classify the processed data. For each sport event class, we randomly select 70 images as a training dataset and 60 images as a testing dataset. We run our proposed algorithm



**Table 2** Average accuracy of different algorithms for scene categorization

Approach	GMM-EM	GMM-RPEM	Probabilistic	SPM	BOW	MLE-Scene	MM-Scene	IAGM
Acc (%)	69.51	69.76	63.88	66.00	71.57	69.87	71.70	73.33



**Fig. 5** Sample frames from UIUC sport event dataset. Our samples show the diversity of background and complexity of information

30 times to obtain the average accuracy results for comparison.

In order to demonstrate the advantages of our algorithm, we compared our model with a number of state-of-the-art approaches within similar area. These approaches include Gaussian mixture model with Expectation Maximization algorithm (GMM-EM) (Law et al. 2004), Gaussian mixture model with Rival Penalized Expectation Maximization (GMM-RPEM) (Cheung and Zeng 2006), GIST (Oliva and Torralba 2001), multi-class supervised Latent Dirichlet Allocation and multi-class supervised Latent Dirichlet Allocation with annotations (probabilistic) (Wang et al. 2009), Spatial pyramid matching (SPM) (Lazebnik et al. 2006), bag of keypoints (BOK) (Csurka et al. 2004), maximum likelihood estimation Scene (MLE-Scene) and Max-Margin Scene (MM-Scene) (Zhu et al. 2010). The evaluation results are shown at Table 2. Fig. 4 displays the confusion matrix for IAGM applied on sport dataset.

We can observe from our results that our proposed IAGM with simultaneous feature selection outperforms other approaches under consideration and provides better average accuracy results for the task of scene categorization.

### 5 Conclusion

In this paper, we present a Dirichlet process mixture model capable of approximating asymmetric Gaussian distributed data, and automatically determining components number, and simultaneously performing feature selection for clustering high-dimensional data. The assumption of asymmetric Gaussian is supported by the fact that natural scene usually not distributed in Gaussian distribution. Dirichlet process allows components number grows to infinite. Infinite mixture model offers flexible representation and straightforward interpretation. Through Bayesian framework, identifying rel-

evant feature and parameter inference are unified into the same framework. We have demonstrated the excellent performance of our algorithm on both dynamic textures clustering and scene categorization tasks.

Although MCMC based Bayesian inference provides a clear posterior sampling approach but it also bring heavier computation cost. A possible future work could be the development of a variational inference based learning approach for the proposed data which is capable of involving massive data and saving tremendous time.

### Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by any of the authors.

### Appendix

Based on the hyperparameters setting chosen in Section 4, we deduce the posteriors for all of the parameters. For parameter  $\alpha$ , the posteriors depend only on the number of observations  $N$  and the number of components  $M$ , and not on how the distributions are distributed among the mixtures:

$$p(\alpha | k, n) \propto \frac{\alpha^{M-\frac{3}{2}} \exp(-\frac{1}{2\alpha}) \Gamma(\alpha)}{\Gamma(N + \alpha)} \tag{32}$$

The complete posteriors for  $\mu$ ,  $\mu_{irr}$ ,  $\lambda$  and  $r$  are obtained as follows:

$$p(\mu_{jk} | \dots) \propto \mathcal{N} \left( (r\lambda + S_{ljk} \sum_{i:\phi_{ijk}=1, X_{ik} < \mu_{jk}} X_{ik} + s_{rjk} \sum_{i:\phi_{ijk}=1, X_{ik} \geq \mu_{jk}} X_{ik}) / (r + pS_{ljk} + (n_j - p)s_{rjk}), \frac{1}{r + pS_{ljk} + (n_j - p)s_{rjk}} \right) \tag{33}$$

$$p(\mu_{jk}^{irr} | \dots) \propto \mathcal{N} \left( \frac{\sum_{i,\phi_{ijk}=0} x_{ik}^{irr} S_{jk}^{irr} + r_k^{irr} \lambda_k^{irr}}{r_k^{irr} + n_j^{irr} S_{jk}^{irr}}, \frac{1}{r_k^{irr} + n_j^{irr} S_{jk}^{irr}} \right) \tag{34}$$

$$p(\lambda \mid \mu_{1k}, \dots, \mu_{Mk}, r) \propto \mathcal{N}\left(\frac{r \sum_{j=1}^M \mu_{jk} + \mu_x \sigma_x^{-2}}{\sigma_x^{-2} + Mr_k}, \frac{1}{\sigma_x^{-2} + Mr_k}\right) \tag{35}$$

$$p(r \mid \mu_{1k}, \dots, \mu_{Mk}, \lambda) \propto \gamma\left(\frac{M+1}{2}, \frac{2}{\sigma_x^2 + \sum_{j=1}^M (\mu_{jk} - \lambda_k)^2}\right) \tag{36}$$

$$\tag{37}$$

The complete posteriors for  $s_{ljk}$ ,  $s_{rjk}$ ,  $s_{jk}^{irr}$ ,  $\beta$  and  $w$  are obtained as follows:

$$p(S_{ljk} \mid X, \mu_j, S_{rj}, \beta_l, w_l) \propto \exp\left[-\frac{S_{ljk} \sum_{i: X_{ik} < \mu_{jk}} (x_{ik} - \mu_{jk})^2}{2} - \frac{w_{lk} \beta_{lk} S_{ljk}}{2}\right] \tag{38}$$

$$p(S_j^{irr} \mid X, \mu_j^{irr}, \beta^{irr}, w^{irr}) \propto \Gamma\left(\frac{N_{jk}^{irr} \beta_k^{irr}}{2}, \frac{\beta_k^{irr} w_k^{irr} + \sum_{i, \phi_{ijk}=0} (X_{ik} - \mu_{jk}^{irr})^2}{2}\right) \tag{39}$$

$$p(\beta_l \mid S_{l1k}, \dots, S_{lMk}, w_l) \propto \Gamma\left(\frac{\beta_l}{2}\right)^{-M} \exp\left(-\frac{1}{2\beta_l}\right) \left(\frac{\beta_l}{2}\right)^{\frac{M\beta_l-3}{2}} \prod_{j=1}^M (w_l S_{ljk})^{\frac{\beta_l}{2}} \exp\left(-\frac{\beta_l w_l S_{ljk}}{2}\right) \tag{40}$$

$$p(w_l \mid S_{l1k}, \dots, S_{lMk}, \beta_l) \propto \Gamma\left(\frac{M\beta_l + 1}{2}, \frac{2}{\sigma_y^{-2} + \beta_l \sum_{j=1}^M S_{ljk}}\right) \tag{41}$$

$N_{jk}^{re}$  and  $N_{jk}^{irr}$  are the number of observations allocated to mixture  $j$  with feature  $k$  considered as relevant and irrelevant, respectively.

The complete posteriors for feature saliency  $\phi$  with gamma parameters  $a$  and  $b$ , with  $n_{jk}$  the number of feature  $k$  relevant for component  $j$  can then be expressed by:

$$p(\rho_{jk} \mid \dots) \propto \text{Beta}(a + n_{jk}, b + N - n_{jk}) \tag{42}$$

$$p(a \mid \dots) \propto a e^{-\frac{a}{2}} \left(\frac{\Gamma(a+b)}{\Gamma(a)}\right)^M \prod_{j=1}^M \rho_{jk}^{a-1}$$

$$p(b \mid \dots) \propto b e^{-\frac{b}{2}} \left(\frac{\Gamma(a+b)}{\Gamma(b)}\right)^M \prod_{j=1}^M (1 - \rho_{jk})^{a-1} \tag{43}$$

### References

Adams S, Beling PA (2017) A survey of feature selection methods for gaussian mixture models and hidden markov models. *Artif Intell Rev* 52:1739–1779

Andonovski G, Mušič G, Blažič S, Škrjanc I (2018) Evolving model identification for process monitoring and prediction of non-linear systems. *Eng Appl Artif Intell* 68:214–221

Antoniak CE (1974) Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Ann Statist* 2(6):1152–1174

Blei DM, Jordan MI (2006) Variational inference for dirichlet process mixtures. *Bayesian Anal* 1(1):121–143

Bouguila N (2009) A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity. *IEEE Trans Knowl Data Eng* 21(12):1649–1664

Bouguila N, Ziou D (2006) Unsupervised selection of a finite dirichlet mixture model: An mml-based approach. *IEEE Trans Knowl Data Eng* 18(8):993–1009

Bouguila N, Ziou D (2012) A countably infinite mixture model for clustering and feature selection. *Knowl Inf Syst* 33(2):351–370

Boutemedjet S, Bouguila N, Ziou D (2009) A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. *IEEE Trans Pattern Anal Mach Intell* 31:1429–1443

Boutemedjet S, Ziou D, Bouguila N (2010) Model-based subspace clustering of non-gaussian data. *Neurocomputing* 73(10–12):1730–1739

Bouveyron C, Brunet-Saumard C (2014) Discriminative variable selection for clustering with the sparse fisher-em algorithm. *Comput Stat* 29(3):489–513

Channoufi I, Bourouis S, Bouguila N, Hamrouni K (2018) Color image segmentation with bounded generalized gaussian mixture model and feature selection. In: 2018 4th International conference on advanced technologies for signal and image processing (ATSIP), pp 1–6

Channoufi I, Bourouis S, Bouguila N, Hamrouni K (2018) Image and video denoising by combining unsupervised bounded generalized gaussian mixture modeling and spatial information. *Multimed Tools Appl* 77(19):25591–25606

Cheung Y, Zeng H (2006) Feature weighted rival penalized em for gaussian mixture clustering: Automatic feature and model selections in a single paradigm. In: 2006 International Conference on Computational Intelligence and Security, vol 1, pp 633–638

Csurka G, Dance CR, Fan L, Willamowski J, Bray C (2004) Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV, pp 1–22

Doretto G, Chiuso A, Wu YN, Soatto S (2003) Dynamic textures. *Int J Comp Vision* 51(2):91–109

Elguebaly T, Bouguila N (2011) Bayesian learning of finite generalized gaussian mixture models on images. *Sig Proces* 91(4):801–820

Elguebaly T, Bouguila N (2012) Generalized gaussian mixture models as a nonparametric bayesian approach for clustering using class-specific visual features. *J Vis Comun Image Represent* 23(8):1199–1212

Elguebaly T, Bouguila N (2014) Background subtraction using finite mixtures of asymmetric gaussian distributions and shadow detection. *Mach Vis Appl* 25(5):1145–1162

Elguebaly T, Bouguila N (2015) Simultaneous high-dimensional clustering and feature selection using asymmetric gaussian mixture models. *Image Vis Comput* 34:27–41

Fan W, Bouguila N (2013) Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, AAAI Press, IJCAI '13, pp 1323–1329

Fan W, Bouguila N (2015) Dynamic textures clustering using a hierarchical pitman-yor process mixture of dirichlet distributions. In:

- 2015 IEEE International Conference on Image Processing (ICIP), pp 296–300
- Fu S, Bouguila N (2018) 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, pp 355–365
- Galimberti G, Manisi A, Soffritti G (2018) Modelling the role of variables in model-based cluster analysis. *Statist Comput* 28(1):145–169
- Griffin JE, Steel MFJ (2010) Bayesian nonparametric modelling with the dirichlet process regression smoother. *Statist Sinica* 20(4):1507–1527
- Gustafson P, Carbonetto P, Thompson N, de Freitas N (2003) Bayesian feature weighting for unsupervised learning, with application to object recognition. In: *AISTATS*
- Hyvärinen A, Hoyer P (2000) Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Comput* 12(7):1705–1720
- Krishnan S, Samudravijaya K, Rao P (1996) Feature selection for pattern classification with gaussian mixture models: a new objective criterion. *Pattern Recognit Lett* 17(8):803–809
- Laptev I (2009) Improving object detection with boosted histograms. *Image Vision Comput* 27(5):535–544
- Law MHC, Figueiredo MAT, Jain AK (2004) Simultaneous feature selection and clustering using mixture models. *IEEE Trans Pattern Anal Mach Intell* 26:1154–1166
- Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol 2, pp 2169–2178
- Li LJ, Fei-Fei L (2007) What, where and who? classifying events by scene and object recognition. In: *2007 IEEE 11th International Conference on Computer Vision* pp 1–8
- Marbac M, Sedki M (2017) Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat Comput* 27(4):1049–1063
- Neal RM (2000) Markov chain sampling methods for dirichlet process mixture models. *J Comput Graph Stat* 9(2):249–265
- Oliva A, Torralba A (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int J Comp Vision* 42(3):145–175
- Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. *J Mach Learn Res* 8:1145–1164
- Park S, Serpedin E, Qaraqe K (2013) Gaussian assumption: The least favorable but the most useful [lecture notes]. *IEEE Signal Process Magaz* 30(3):183–186
- Péteri R, Fazekas S, Huiskes MJ (2010) DynTex: a Comprehensive Database of Dynamic Textures. *Pattern Recognit Lett* 31:1627–1632
- Rasmussen CE (1999) The infinite gaussian mixture model. In: *Proceedings of the 12th International Conference on Neural Information Processing Systems*, MIT Press, Cambridge, MA, USA, NIPS'99, pp 554–560
- Škrjanc I, Iglesias JA, Sanchis A, Leite D, Lughofer E, Gomide F (2019) Evolving fuzzy and neuro-fuzzy approaches in clustering, regression, identification, and classification: A survey. *Inf Sci* 490:344–368
- Song Z, Ali S, Bouguila N (2019) Bayesian learning of infinite asymmetric gaussian mixture models for background subtraction. In: *Image Analysis and Recognition—16th International Conference, ICIAR 2019*, Waterloo, ON, Canada, August 27–29, 2019, Proceedings, Part I, pp 264–274
- Wang C, Blei DM, Fei-Fei L (2009) Simultaneous image classification and annotation. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* pp 1903–1910
- Wang S, Zhu J (2008) Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* 64(2):440–8
- Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans Pattern Anal Mach Intell* 29(6):915–928
- Zhu J, Li LJ, Fei-Fei L, Xing EP (2010) Large margin learning of upstream scene understanding models. In: *NIPS*

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.