Bayesian Non-Parametric Parsimonious Gaussian Mixture for Clustering

Faicel Chamroukhi*†, Marius Bartcus*† and Hervé Glotin *†‡
* Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
†Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
‡Institut Universitaire de France (IUF), 75005 Paris
Email: {chamroukhi, bartcus, glotin}@univ-tln.fr

***Abstract*—Clustering is one of the essential tasks in machine learning and statistical pattern recognition. One of the most popular approaches in cluster analysis is the one based on the parametric finite mixture model. However, often, parametric models are not well adapted to represent complex and realistic data sets. Another issue in the finite mixture model-based clustering approach is the one of selecting the number of mixture components. The Bayesian non-parametric statistical methods for clustering provide a principled way to overcome these issues. This paper proposes a new Bayesian non-parametric approach for clustering. It relies on an Infinite Gaussian mixture model with an eigenvalue decomposition of the covariance matrix of each cluster, and a Chinese Restaurant Process (CRP) prior over the hidden partition. The CRP prior allows to control the model complexity in a principled way, and to automatically learn the number of clusters from the data. The covariance matrix decomposition allows to fit various flexible models going from simplest spherical ones to the more complex general one. We develop a Gibbs sampler to learn the various models and apply it to simulated data and benchmarks, and a real-world data issued from a challenging problem of whale song decomposition. The obtained results highlight the interest of the proposed non-parametric parsimonious mixture model for clustering.**

I. INTRODUCTION

Clustering is one of the essential tasks in machine learning and statistics. One of the most popular approaches in cluster analysis is the one based on the parametric finite mixture model [1][2]. However, these parametric models may not be well adapted to represent complex and realistic data sets. Another issue in the finite mixture model-based clustering approach is the one of selecting the number of mixtures (model selection). Bayesian Non-Parametric (BNP) [3][4] methods for clustering, including Infinite Gaussian Mixture Models (IGMM) [5], Chinese Restaurant Process (CRP) mixtures and Dirichlet Process Mixtures (DPM) [6][7][8][9][10] provide a principled way to overcome these issues. They avoid assuming restricted functional forms and thus allow the complexity and accuracy of the inferred model to grow as more data is observed. They also represent a good alternative to the difficult problem of model selection in parametric mixture models, namely the finite Gaussain Mixture Model (GMM).

In this work, we rely on the BNP formulation of the Gaussian mixture and a flexible decomposition of the covariance matrix of each Gaussian density which has proven its big flexibility in cluster analysis [11][12][2]. This leads to an Infinite Parsimonious Gaussian Mixture Model (IPGMM) which is more flexible in term of modeling and its use in clustering, and automatically provides the number of clusters.

The paper is organized as follows: Section II briefly discusses previous work on finite Gaussian mixture clustering. Then, section III presents the proposed model and its learning technique. In section IV, we give experimental results to evaluate the proposed approach.

II. PARAMETRIC PARSIMONIOUS GAUSSIAN CLUSTERING

A. Model

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a sample of n i.i.d observations in $\mathbb{R}^d$, and let $\mathbf{z} = (z_1, \ldots, z_n)$ be the corresponding unknown cluster labels where $z_i \in \{1, \ldots, K\}$ represents the cluster label of the ith data point $\mathbf{x}_i$, K being the possibly unknown number of clusters.

Parametric Gaussian clustering, also called model-based clustering [13] [2], is based on the finite GMM [1] in which the probability density function of the data is given by:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \ \mathcal{N}_k(\mathbf{x}_i|\theta_k) \tag{1}$$

where $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1}^K$ are the GMM parameters which include the non-negative mixing proportions π_k that sum to one and $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ which are respectively the mean vector and the covariance matrix for the kth Gaussian component density.

The GMM clustering has been extended to parsimonious GMM clustering [12][11] by exploiting an eigenvalue decomposition of the group covariance matrices, which provides a wide range of very flexible models with different clustering criteria. The eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}_k$ of each cluster k is given by:

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{2}$$

where $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/d}$, $\mathbf{D}_k$ is an orthogonal matrix of eigenvectors of $\boldsymbol{\Sigma}_k$ and $\mathbf{A}_k$ is a diagonal matrix with determinant 1 whose diagonal elements are the normalized eigenvalues of $\boldsymbol{\Sigma}_k$ in a decreasing order. As pointed in [12], the scalar λ_k determines the volume of cluster k, $\mathbf{D}_k$ its orientation and $\mathbf{A}_k$ its shape. Thus, this decomposition leads to fourteen flexible models [12] going from simplest spherical models to the complex general one ans hence is adapted to various clustering situations.

B. Parameter estimation

The parameters $\boldsymbol{\theta}$ of the Gaussian mixture can be estimated in a maximum likelihood (ML) framework by maximizing the observed data likelihood

$$p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\sum_{k=1}^{K} \pi_k \; \mathcal{N}_k(\mathbf{x}_i|\theta_k), \tag{3}$$

or in a maximum a posteriori (MAP) estimation (Bayesian) framework by maximizing the posterior parameter distribution

$$p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta}), \tag{4}$$

$p(\boldsymbol{\theta})$ being a prior distribution on the model parameters $\boldsymbol{\theta}$. The maximum likelihood estimation usually relies on the Expectation-Maximization (EM) algorithm [14][15] or EM extensions. The maximization of the posterior can still be performed by EM, namely in the case of conjugate priors as in [16]. It can also be performed by Markov Chain Monte Carlo (MCMC) sampling techniques as in [17][18][19]. For the case of parsimonious finite Gaussian mixture, several learning algorithms have also been proposed. They in majority rely on a maximum likelihood estimation via EM or EM extensions [11][12] or on Bayesian (MAP) estimation using EM as in [16] or by MCMC sampling techniques like the Gibbs sampler as in [17][18].

However, in the finite GMM approach for clustering, the number of clusters is required. One of the main issues in the parametric model-based clustering is therefore the one of selecting the number of mixture components (clusters) that fit at best the data. The choice of the optimal number of clusters can be performed via penalized log-likelihood criteria such as the Bayesian Information Criterion (BIC)[20] or the Akaike Information Criterion (AIC) [21], etc.

BNP mixtures for clustering offer a principled alternative to infer the number of clusters from the data in a single run, rather than in a two-stage approach as in standard model-based clustering [4][6][5]. They assume that the observed data are governed by an infinite number of clusters, but only a finite number of them does actually generates the data. In the next section, we rely on the infinite mixture model formulation to derive the proposed approach.

III. The proposed Bayesian non-parametric parsimonious Gaussian mixture for clustering

BNP mixture approaches for clustering assume general process as prior on the infinite possible partitions, which is not restrictive as in classical Bayesian inference. Such a prior can be a DP [10][9][6] or CRP [22][6]. Several BNP models have considered the general GMM, that is the infinite GMM [5] which can have interpretation in term of the CRP mixture [6] or by equivalence the DPM [9][6]. For additional review on BNP clustering, see for example [6].

The proposed BNP parsimonious clustering approach exploits the eigenvalue decomposition of the cluster covariance matrices as in [12][11] and integrates it into an infinite mixture modeling framework by using a CRP prior. This leads to an Infinite Parsimonious Gaussian Mixture Model (IPGMM) which is very flexible in terms of modeling, and automatically infers the optimal number clusters from the data. In the next section, we derive the proposed CRP mixture in the case of the parsimonious model, and then we provide an MCMC estimation technique for the derived models.

A. Chinese Restaurant Process (CRP) parsimonious mixture

The CRP provides a distribution on the infinite partitions of the data, that is a distribution over the positive integers $1, \ldots, n$. Consider the following joint distribution of the unknown cluster assignments:

$$p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1)\ldots p(z_n|z_1, z_2, \ldots, z_{n-1})\cdot \tag{5}$$

Each term of this joint distribution can be computed from the CRP prior as follows. Suppose there is a restaurant with an infinite number of tables and in which customers are entering and sitting at tables. We assume that customers are social, so that the ith customer sits at table k with probability proportional to the number of already seated customers n_k, and may choose a new table with a probability proportional to a small positive real number α which represents the CRP concentration parameter. This can be explicitly formulated as follows

$$\begin{aligned} p(z_i = k|z_1, ..., z_{i-1}) &= \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \\ &= \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_+ \end{cases} \end{aligned} \tag{6}$$

where K_+ is the number of tables for which the number of customers sitting in $n_k > 0$, and for $k \leq K_+$ that means that k is a previously unoccupied table and for $k > K_+$ that means a new table to be occupied. From this distribution, one can therefore allow assigning new data to possibly previously unseen (new) clusters as the data are observed, after starting with one cluster. In clustering with the CRP, customers correspond to data points and tables correspond to clusters. In CRP mixture, the prior $\text{CRP}(z_1, \ldots, z_{i-1}; \alpha)$ is completed with a likelihood with parameters $\boldsymbol{\theta}_k$ with each table (cluster) k (i.e., a multivariate Gaussian likelihood with mean vector and covariance matrix in the GMM case), and a prior distribution (G_0) for the parameters. For example, in the GMM case, one can use a conjugate multivariate normal inverse-Wishart prior distribution for the mean vectors and the covariance matrices. This corresponds to the ith customer sits at table $z_i = k$ chooses a dish (the parameter $\boldsymbol{\theta}_{z_i}$) from the prior of that table (cluster). The CRP mixture can be summarized according to the following generative process.

$$z_i \sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \tag{7}$$

$$\boldsymbol{\theta}_{z_i} \sim G_0 \tag{8}$$

$$\mathbf{x}_i \sim p(.|\boldsymbol{\theta}_{z_i})\cdot \tag{9}$$

In our proposed infinite parsimonious Gaussian mixture, the cluster covariance matrices are parametrized in term of an eigenvalue decomposition to provide more flexible clusters with possibly different volumes, shapes and orientations. This can be seen as a variability of dishes in terms of Chinese Restaurant interpretation. Note that one can also give interpretation of the CRP mixture in terms of DPM [10][6].

B. MCMC Gibbs sampling for model learning

We developed an MCMC Gibbs sampling technique, as in [5][23][8][6], to learn the proposed Bayesian non-parametric

Decomposition	Model-Type	Prior	Applied to
$\lambda\mathbf{I}$	Spherical	$\mathcal{IG}$	λ
$\lambda_k\mathbf{I}$	Spherical	$\mathcal{IG}$	λ_k
$\lambda\mathbf{A}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
$\lambda_k\mathbf{A}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}$
$\lambda\mathbf{DAD}^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

TABLE I. CONSIDERED PARSIMONIOUS GMMS VIA EIGENVALUE DECOMPOSITION AND THE ASSOCIATED PRIOR FOR THE COVARIANCE. NOTE THAT $\mathcal{I}$ DENOTES AN INVERSE DISTRIBUTION, $\mathcal{G}$ A GAMMA DISTRIBUTION AND $\mathcal{W}$ A WISHART DISTRIBUTION.

parsimonious mixture model. The developed Gibbs sampler is summarized by the pseudo-code (1). The used priors on the model parameters depend on the type of the parsimonious model (see Table I). Thus, sampling the model parameters varies according to the considered parsimonious mixture model. Indeed, yet we investigated seven parsimonious models, covering the three families of the mixture models: the general, the diagonal and the spherical family. The parsimonious models therefore go from the simplest spherical one to the more general full model. Table I summarizes the considered models and the corresponding prior for each model used in Gibbs sampling. We note that the resulting posterior distributions for the considered models are close to those in [17].

Algorithm 1 Gibbs sampling for the proposed IPGMM

```
Inputs: a data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, hyper-parameters and number of Gibbs samples
Initialize the model hyper-parameters $H$.
Start with one cluster $K_+ = 1, \boldsymbol{\theta}_1 = \{\boldsymbol{\mu}_1, \mathbf{\Sigma}_1\}$
for $t = 2, \ldots, \#$samples do
    for $i = 1, \ldots, n$ do
        for $k = 1, \ldots, K_+$ do
            if $(n_k = \sum_{i=1}^{N} z_{ik}) - 1 = 0$ then
                We decrease $K_+ = K_+ - 1$; $\boldsymbol{\theta}^{(t)} = \{\boldsymbol{\theta}^{(t)}\} \setminus \boldsymbol{\theta}_{z_i}$
            end if
        end for
        Sample a cluster label $z_i^{(t)}$ from the posterior:
        $p\left(z_i|\mathbf{z}_{\backslash z_i}, \mathbf{X}, \boldsymbol{\theta}^{(t)}, H\right) \propto p\left(\mathbf{x}_i|z_i, \boldsymbol{\theta}^{(t)}\right)$ CRP$(\mathbf{z}_{\backslash z_i}; \alpha)$
        if $z_i^{(t)} = K_+ + 1$ then
            We get a new cluster, we increase $K_+ = K_+ + 1$ and we sample a new cluster parameter $\boldsymbol{\theta}_{z_i}^{(t)}$ from the prior distribution as in Table I
        end if
    end for
    for $i = 1, \ldots, K_+$ do
        Sample the parameters $\boldsymbol{\theta}_k^{(t)}$ from the posterior distribution.
    end for
    Sample the hyperparameter $\alpha^{(t)} \sim p(\alpha^{(t)}|K_+) \propto \mathcal{G}(a, b)$ [24]
    $\mathbf{z}^{(t+1)} \leftarrow \mathbf{z}^{(t)}$
end for
Outputs: $\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}, \hat{K} = K_+\}$
```

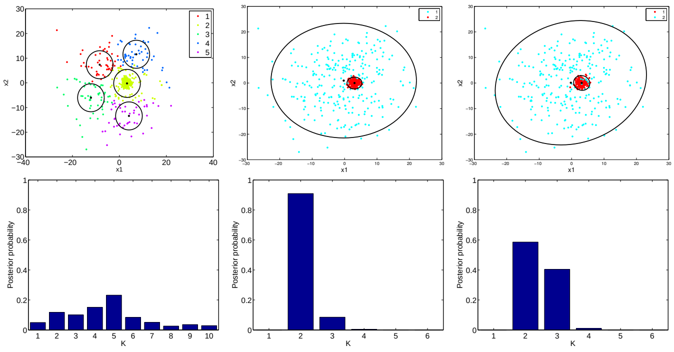
IV. EXPERIMENTS

We performed experiments on both simulated and real data in order to test our proposed non-parametric method. We highlight its flexibility in terms of modeling and its use for clustering, as well as inferring the number of clusters from the data. The numerical results are reported in terms of comparisons of the observed-data log-likelihood, the estimated partition of the data, and the selection of the actual number of clusters, for different candidate models. When the number of clusters for the estimated partition equals the actual one, we also report the misclassification error rate. We compared our Bayesian non-parametric parsimonious mixture with different alternatives including: model-based clustering and Bayesian parametric clustering approaches using finite Gaussian mixtures. In the experiments, each algorithm is run ten times with different initializations and the Gibbs sampler generates 2000 samples where the first 200 samples was discarded as burn-in. The best solution, corresponding to the highest posterior probability is then selected.

A. Experiment on simulated data and benchmarks

1) Experiment on simulated data: We first considered a two-class situation to illustrate the interest of the proposed clustering approach. This situation is the same as for the parametric parsimonious mixture approach proposed in [12]. It consists in a sample of $n = 500$ observations from a two-component Gaussian mixture in $\mathbb{R}^2$ with the following parameters: $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (0, 0)^T$ and $\boldsymbol{\mu}_2 = (3, 0)^T$, $\mathbf{\Sigma}_1 = 100\ \mathbf{I}_2$ and $\mathbf{\Sigma}_2 = \mathbf{I}_2$.

Figure 1 shows the simulated data and the obtained partitions by the proposed Bayesian non-parametric clustering approach for three different parsimonious models, and the posterior distribution of the number of clusters K for each model. Table II reports the estimated number of clusters and the



$\lambda\mathbf{I}$ $\quad$ $\lambda_k\mathbf{I}$ $\quad$ $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

$\hat{K} = 5, \hat{\mathcal{L}} = -5900.3$ $\quad$ $\hat{K} = 2, \hat{\mathcal{L}} = -\mathbf{5112.5}$ $\quad$ $\hat{K} = 2, \hat{\mathcal{L}} = -5119.4$

Fig. 1. A two-class data set with (top): the log-likelihood values ($\hat{\mathcal{L}}$) and estimated number of clusters ($\hat{K}$) obtained by three proposed IPGMM: spherical model with identical cluster volumes (left), spherical model with different cluster volumes (middle) and general model (right), and (bottom): the posterior distribution of the number of clusters.

obtained values of the log-likelihoods for each corresponding partition obtained by the parametric Bayesian approach based on Gaussian finite mixture (GMM) and the proposed non-parametric approach (IPGMM). First, it can be observed that, the partition provided by the spherical model ($\lambda\mathbf{I}$) which does not allow clusters with different volumes, can not reconstruct the actual partition. This model also fails for the finite GMM case [12]. However, the spherical model $\lambda_k\mathbf{I}$, which allows different cluster volumes, fits at best the underlying structure of the data and provides a precise partition (the err-rate equals 4.80%) with the actual number of clusters. It is even slightly more precise than the general model. Indeed, the general model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$, which is the more complex model in terms the number of parameters, provides a closely similar result (the err-rate equals 4.40%). From the posterior cluster distributions, we can also see that the spherical model with different cluster volumes ($\lambda_k\mathbf{I}$) is the model that reveals at best the actual number of clusters compared to the other models. On the other hand, we note that the diagonal models can not provide an accurate partition, even the one allowing different volumes ($\lambda_k\mathbf{A}$). Furthermore, for this simulated data, the best log-likelihood value, as it can be seen in Table II, corresponds to the spherical model with different cluster volumes ($\lambda_k\mathbf{I}$). One can conclude that, in a non-parametric clustering, it is important to consider clusters with different volumes, and at least for this data set, the spherical model with different cluster volumes ($\lambda_k\mathbf{I}$) is the best model.

	GMM		IPGMM	
Model	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda\mathbf{I}$	2	-5583.6	5	-5900.3
$\lambda_k\mathbf{I}$	5	-5157.7	2	$\mathbf{-5112.5}$
$\lambda\mathbf{A}$	4	-5474.5	9	-5428.9
$\lambda_k\mathbf{A}$	5	-5157.7	7	-5288.8
$\lambda\mathbf{DAD}^T$	4	-5417.5	7	-5312.7
$\lambda_k\mathbf{DAD}^T$	2	-5560.8	8	-5412.5
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	5	$\mathbf{-5.0938}$	2	-5.1194

TABLE II. LOG-LIKELIHOOD VALUES AND THE ESTIMATED NUMBER OF CLUSTERS OBTAINED BY THE FINITE GMM AND THE PROPOSED IPGMM FOR THE SIMULATED DATA.

2) Experiment on benchmarks: In this experiment, we considered well-known real data sets[1] of Iris, Old Faithful Geyser, Trees, Wine and Diabetes. Table III shows the number of observations, the dimension and the (possibly known) number of clusters for each dataset.

Dataset	Num. of data items (n)	dimension (d)	True K
Iris	150	4	3
Old Faithful Geyser	272	2	Unknown
Trees	31	3	Unknown
Wine	178	13	3
Diabetes	145	3	3

TABLE III. DESCRIPTION OF THE USED BENCHMARKS DATA.

Figure 2 shows the partition and densities estimated by the proposed non-parametric parsimonious clustering approach for Iris (top) and Geyser (bottom) with a spherical model (left), a diagonal model (middle) and the general model (right). The three models allow different cluster volumes. It also shows the posterior distribution of the number of clusters for each model for each of the two data sets. We note, that while the number of classes for Geyser dataset is unknown, several clustering studies estimate it between two and three. Table IV reports the numerical results for the two datasets. We can see on Figure 2 that, for Iris data, both the spherical model $\lambda_k\mathbf{I}$ and the diagonal model $\lambda_k\mathbf{A}$ which consider different cluster volume provide the correct number of classes and allow to reconstruct the hidden data structures. The misclassification error rate for the diagonal model is 5.33% and the one for the spherical model is 10.66%. However, the general model underestimate the number of clusters. Let us also note that, for the finite GMM clustering approach, the models which provide the correct number of clusters are the diagonal models $\lambda\mathbf{A}$ and $\lambda_k\mathbf{A}$ and the corresponding misclassification error rates are respectively 9.33% and 11.33%. This can make more advantageous our non-parametric alternative. The general model provides two clusters for this dataset. Table IV also shows that, while the general model provides two clusters for both the GMM approach and the proposed IPGGM one, the partition provided by the proposed approach is the more likely (best like-likelihood).

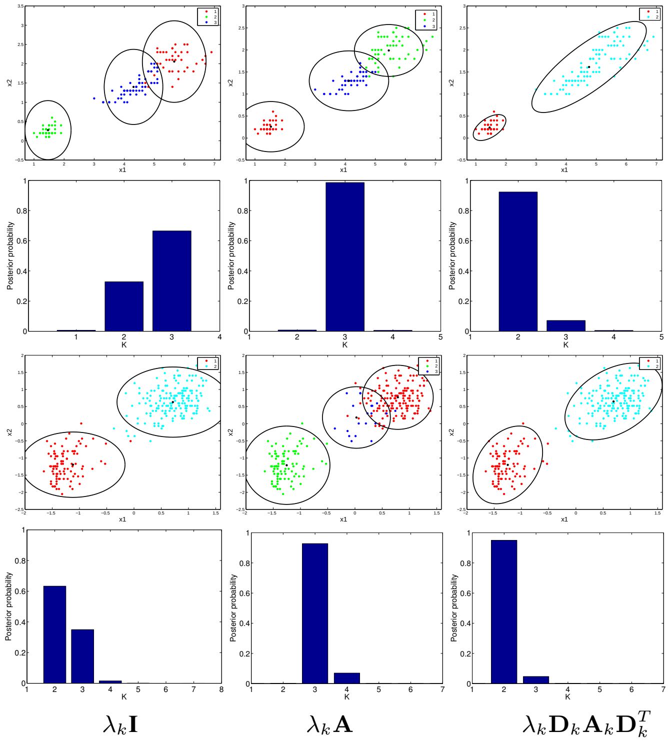


Fig. 2. Clustering results for Iris (top) and Geyser (bottom) obtained by a spherical model (left), a diagonal model (middle) and the general model (right), all allowing different cluster volumes, and the corresponding posterior distributions of the number of clusters.

For the Geyser data, it can be observed on the graphical results that the partitions obtained by the spherical model with different volumes ($\lambda_k\mathbf{I}$) and the general model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ are similar and can be seen as valid. The partition provided by the diagonal model $\lambda_k\mathbf{A}$ can also be accepted. We also note that, except for the spherical model with equal volumes ($\lambda\mathbf{I}$), the estimated number of classes for the other models is consistent with the literature (between two and three). Note that for this data set, the parametric approach based on finite GMMs provides likewise two clusters except for the spherical model with equal cluster volumes. The spherical model with

[1] from the UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/index.html

	Iris				Geyser			
	GMM		IPGMM		GMM		IPGMM	
True value of K	3				-			
Model	$\hat{K}$	log-lik	$\hat{K}$	log-lik	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda\mathbf{I}$	5	-1643.5	5	-1712.6	3	-1597.4	10	-1659.8
$\lambda_k\mathbf{I}$	5	-1663.8	3	-1722.8	2	-1630.6	2	-1634.5
$\lambda\mathbf{A}$	3	-1700.4	4	-1647.5	2	-1622.2	3	-1605.1
$\lambda_k\mathbf{A}$	3	-1714.7	3	-1707.9	2	-1639.6	3	-1609.0
$\lambda\mathbf{DAD}^T$	2	-1629.3	4	-1562.8	2	-1638.0	2	-1601.9
$\lambda_k\mathbf{DAD}^T$	2	-1641.6	4	-1566.4	2	-1605.6	3	$\mathbf{-1593.9}$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	2	$\mathbf{-1583.1}$	2	$\mathbf{-1559.7}$	2	$\mathbf{-1594.7}$	2	-1595.7

TABLE IV. LOG-LIKELIHOOD VALUES AND THE ESTIMATED NUMBER OF CLUSTERS OBTAINED BY THE PARAMETRIC MIXTURE (GMM) AND THE PROPOSED BAYESIAN NON-PARAMETRIC PARSIMONIOUS MIXTURE (IPGMM) FOR IRIS DATA AND GEYSER DATA.

different volumes ($\lambda_k\mathbf{I}$) and the two more likely models are the two general models $\lambda_k\mathbf{DAD}^T$ and $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ for a partition into two and three clusters, both can be selected as valid partitions. Finally, Table V shows the estimated number of clusters for each of the four datasets. For the Trees data, we observe that the majority of the parsimonious models in our Bayesian non-parametric approach estimate 2 clusters. We note that these results are similar to those obtained by[16] who provide clustering results using EM with ML estimation and with MAP estimation using a parametric Bayesian Gaussian mixture. They indeed obtain two clusters when using a prior, and three and five classes without using a prior. For the Diabetes dataset we can see that most of our parsimonious models automatically infer the good number of clusters (three), similarly as in the model-based clustering approach [11]. For the Wine dataset too, the retrieved number of clusters, for the majority of models equals the actual one.

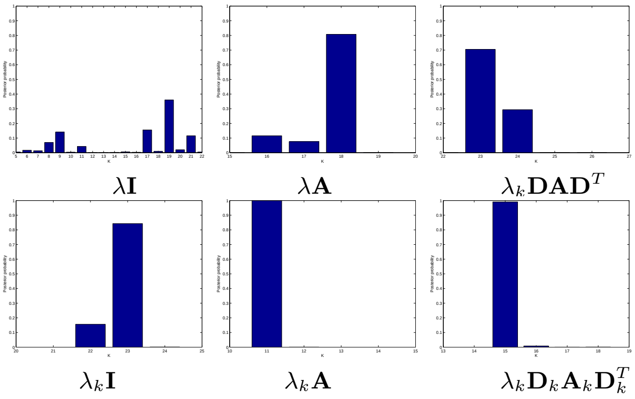
Model	Iris	Geyser	Trees	Wine	Diabetes
$\lambda\mathbf{I}$	5	10	1	1	3
$\lambda_k\mathbf{I}$	3	2	1	2	5
$\lambda\mathbf{A}$	4	3	2	3	3
$\lambda_k\mathbf{A}$	3	3	2	1	5
$\lambda\mathbf{DAD}^T$	4	2	2	3	5
$\lambda_k\mathbf{DAD}^T$	4	3	2	3	3
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	2	2	2	3	3

TABLE V. THE NUMBER OF CLUSTERS PROVIDED BY THE PROPOSED IPGMM.

B. Experiment on real data: whale song decomposition

In this experiment, we apply the proposed approach to a challenging problem of humpback whale song decomposition. Humpback whales produce songs with a specific structure and the study of that songs is very challenging and very useful for bio-acousticians and scientists to namely understand how do whales song and communicate (possibly according to which vocabulary) and to have an idea about their origin, since the songs of whales from different origins can be different. The analysis of such complex signals that aims at discovering the call units (which can be considered as a kind of whale vocabulary), can be seen as a problem of unsupervised call units classification as in [25]. We therefore reformulate the problem of whale song decomposition as a clustering problem. Contrary to the approach used in [25], in which the number of clusters (call units in this case) has been fixed manually, here, we apply our proposed IPGMM to find a partition of the whale song into clusters, and automatically infer the number of clusters from the data. The used data are available in the framework of our SABIOD project[2]. The data consist of MFCC parameters of 8.6 minutes of a Humpback whale song recordings produced at few meters distance from the whale in La Reunion - Indian Ocean.

Figure 3 shows the posterior distribution of the number of clusters for different Bayesian non-parametric parsimonious mixtures for the considered whale song signals. Table VI shows the number of estimated clusters and the log-likelihood values obtained by the parametric GMM approach and the proposed Bayesian non-parametric parsimonious method for clustering the whale song data.


$\lambda\mathbf{I}$ $\lambda\mathbf{A}$ $\lambda_k\mathbf{DAD}^T$

$\lambda_k\mathbf{I}$ $\lambda_k\mathbf{A}$ $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Fig. 3. Posterior distribution of the number of clusters of the whale song data obtained by the proposed Bayesian non-parametric approach.

	EM ML		IPGMM	
Model	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda\mathbf{I}$	60	-2219.8	9	-2341.3
$\lambda_k\mathbf{I}$	60	-2112.9	23	-2213.3
$\lambda\mathbf{A}$	22	-2143.5	18	-2195.8
$\lambda_k\mathbf{A}$	59	-2005.9	11	-2190.0
$\lambda_k\mathbf{DAD}^T$	51	-1981.1	24	-2158.9
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	19	-1.9418	15	-2.1234

TABLE VI. LOG-LIKELIHOOD VALUES (DIVIDED BY 10^3) AND THE NUMBER OF ESTIMATED CLUSTERS FOR THE WHALE DATA.

One can see that the parametric approach in the majority of cases seems to overestimate the number of whale songs, because for this Humpback whale specie, in previous studies, namely in [25], the experts estimated (manually) the number of clusters of about 18 clusters with sometimes three additional clusters, that is 21 clusters. On the other hand, the models of the proposed non-parametric approach provide a plausible number of whale song units. They seem to cover the assumed number of clusters (18, 21), even if no ground truth is available

[2] Scaled Acoustic BIODiversity: http://sabiod.univ-tln.fr/data_samples.html

for this specific data set. The two general models $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ and $\lambda_k \mathbf{DAD}^T$, which are the more likely models for this whale song dataset, provide respectively 15 and 24 song units which are also reasonable. However, the simple spherical model $\lambda \mathbf{I}$ seems to be not adapted for this task. Now, we analyze the result provided by the model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ and we show in Figure 4 the whale song partition, which correspond to the whale song decomposition into several units. From this decomposition, the clusters 8, 12 and 15 are uniformly activated in time, therefore they would correspond to the background (sea) noise, rather than actual whale song units. Whereas, the remaining clusters are likely to correspond to the whale song units. This can be more observed on Figure 5 which shows the spectrograms of the whale song units from the partition obtained with the proposed approach and on which we can see that the obtained song units are clearly conveying information.

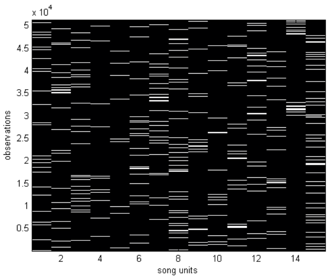


Fig. 4. Clustering partition of the whale song obtained by the non-parametric approach (IPGMM) with the model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$.

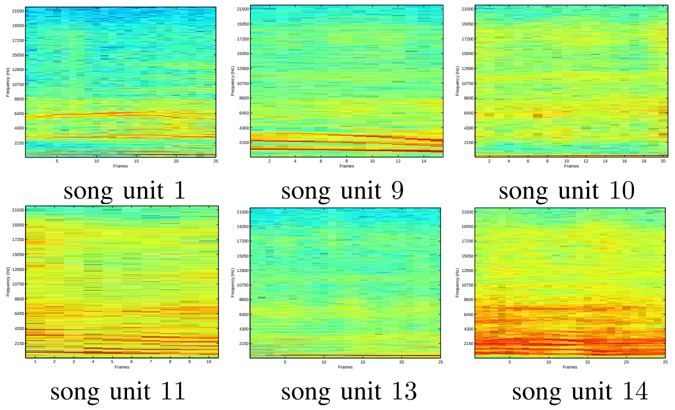


Fig. 5. Whale songs spectrograms obtained with the proposed Bayesian non-parametric approach with the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$.

V. Conclusion

In this paper we presented a new Bayesian non-parametric parsimonious mixture approach for clustering. It is based on an infinite Gaussian mixture with an eigenvalue decomposition of the cluster covariance matrix and a Chinese Restaurant Process prior. It allows deriving several flexible models and avoids the problem of model selection encountered in the standard maximum likelihood-based and Bayesian parametric Gaussian mixture. We illustrated this method on simulated data and benchmarks, and applied it to a challenging problem of clustering bio-acoustic data. The obtaining results highlight the interest of using parsimonious Bayesian clustering as a good alternative to finite GMM clustering. Future work will concern namely Bayesian model comparison.

Acknowledgment

This work is supported by the SABIOD project: http://sabiod.univ-tln.fr and PHRASE DGA RAPID.

References

[1] G. J. McLachlan and D. Peel., *Finite mixture models*. New York: Wiley, 2000.

[2] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[3] C. P. Robert, *The Bayesian choice: a decision-theoretic motivation*. Springer-Verlag, 1994.

[4] N. Hjort, C. Holmes, P. Muller, and S. G. Waller, *Bayesian Non Parametrics: Principles and practice*, C. U. Press, Ed., 2010.

[5] C. Rasmussen, "The infinite gaussian mixture model." *Advances in neuronal Information Processing Systems*, vol. 10, pp. 554 – 560, 2000.

[6] J. G. Samuel and D. M. Blei, "A tutorial on bayesian non-parametric model," *Journal of Mathematical Psychology*, vol. 56, pp. 1–12, 2012.

[7] J. Pitman, "Exchangeable and partially exchangeable random partitions," *Prob. Theory Related Fields*, vol. 102, no. 2, pp. 145–158, 1995.

[8] F. Wood, T. L. Griffiths, and Z. Ghahramani, "A non-parametric bayesian method for inferring hidden causes." in *UAI*, 2006.

[9] C. E. Antoniak, "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[10] T. S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[11] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, no. 3, pp. 803–821, 1993.

[12] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models." *Pattern Recognition*, vol. 28, no. 5, pp. 781–793, 1995.

[13] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988.

[14] 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, B*, vol. 39(1), pp. 1–38, 1977.

[15] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. New York: Wiley, 1997.

[16] C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2, pp. 155–181, Sep. 2007.

[17] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert, "Inference in model-based cluster analysis," *Statistics and Computing*, vol. 7, no. 1, pp. 1–10, 1997.

[18] H. Bensmail and J. J. Meulman, "Model-based clustering with noise: Bayesian inference and estimation," *Journal of Classification*, vol. 20, no. 1, pp. 049–076, 2003.

[19] D. Ormoneit and V. Tresp, "Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates." *IEEE Transactions on Neural Networks*, vol. 9, no. 4, pp. 639–650, 1998.

[20] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

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

[22] J. Pitman, "Combinatorial stochastic processes," Dept. of Statistics. UC, Berkeley, Tech. Rep. 621, 2002.

[23] R. M. Neal, "Probabilistic inference using markov chain monte carlo methods," Dept. of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993.

[24] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 577–588, 1994.

[25] F. Pace, F. Benard, H. Glotin, O. Adam, and P. White, "Subunit definition and analysis for humpback whale call classification," *Applied Acoustics*, vol. 71, no. 11, pp. 1107 – 1112, 2010.