# Bayesian Non-Parametric Parsimonious Clustering

Faicel Chamroukhi[1,2], Marius Bartcus[1,2], Hervé Glotin[1,2,3]

1- Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
2- Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
3- Institut Universitaire de France, iuf.amue.fr

**Abstract**. This paper proposes a new Bayesian non-parametric approach for clustering. It relies on an infinite Gaussian mixture model with a Chinese Restaurant Process (CRP) prior, and an eigenvalue decomposition of the covariance matrix of each cluster. The CRP prior allows to control the model complexity in a principled way and to automatically learn the number of clusters. The covariance matrix decomposition allows to fit various parsimonious models going from simplest spherical ones to the more complex general one. We develop an MCMC Gibbs sampler to learn the models. First results obtained on both simulated and real data highlight the interest of the proposed infinite parsimonious mixture model.

## 1 Introduction

Clustering is one of the essential tasks in machine learning and statistics. One of the most popular approaches in cluster analysis is the parametric finite mixture model-based clustering [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) methods for clustering, including Infinite Gaussian Mixture Models (IGMM) [3] and CRP mixtures [4] provide a principled way to overcome these issues. They avoid the assumption of restricted functional forms and thus allow the complexity and accuracy of the inferred model to grow as more data is observed. The non-parametric aspect of these approaches relates the hypothesis of assuming that model complexity associated to the number of model parameters grows with the data volume and complexity. These non-parametric approaches also represent a good alternative to the difficult problem of model selection encountered in parametric models. In this work , we rely on this Bayesian non-parametric formulation of the GMM and assume the flexible decomposition of the covariance matrix of each Gaussian density which has proven its big flexibility in cluster analysis [5, 6]. This leads to an Infinite Parsimonious Gaussian mixture 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 2 briefly discusses previous work on finite Gaussian mixture clustering. Then, Section 3 presents the proposed approach and Section 4 shows experiment results.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a set of $n$ i.i.d multidimensional data in $\mathbb{R}^d$, and let $\mathbf{z} = (z_1, \ldots, z_n)$ be the corresponding unknown cluster labels where $z_i \in \{1, \ldots, K\}$,

## 2 Parametric parsimonious Gaussian clustering

Parametric Gaussian clustering is based on the finite Gaussian Mixture Model (GMM) [1, 2] where 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 $\pi_k$ 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 $k$th Gaussian component. The finite parsimonious GMMs [5, 6] exploit an eigenvalue decomposition of the Gaussian covariance matrices. This provides a wide range of very flexible models going from simplest spherical models to the complex general one. Indeed, the eigenvalue decomposition of the covariance matrix for each Gaussian component density allows having clusters with diffentes volumes, orientations and shapes [5, 6]. This parametrization of the covariance matrix is of the following form:

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

where $\lambda_k$ is a scalar that defines the volume of cluster $k$, $\mathbf{D}_k$ is an orthogonal matrix which defines its orientation and $\mathbf{A}_k$ is a diagonal matrix with determinant 1 which defines its shape. The mixture model parameters $\boldsymbol{\theta}$ can be estimated by maximizing the observed data likelihood $p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \; \mathcal{N}_k(\mathbf{x}_i|\theta_k)$ or in a maximum a posteriori (MAP) estimation (Bayesian) framework by maximizing the posterior parameter distribution: $p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ being a chosen prior distribution on the model parameters $\boldsymbol{\theta}$. The Maximum Likelihood estimation usually relies on the Expectation-Maximization (EM) algorithm or EM extensions [7]. The MAP estimation can still be performed by EM in the case of conjugate priors as in [8]. Markov Chain Monte Carlo (MCMC) sampling techniques can also be used as in [9]. For the case of finite parsimonious GMMs, several learning algorithms have been proposed. They in majority rely on a ML estimation via EM or EM extensions [5, 6], or on Bayesian (MAP) estimation using EM as in [8] or by MCMC sampling techniques, namely the Gibbs sampler as in [9]. However, in the finite GMM approach for clustering, the number of clusters is required. One of the main issues in parametric model-based clustering is therefore the one of selecting the number of clusters. This model selection in parametric Bayesian and non-Bayesian mixture clustering can be performed via penalized log-likelihood criteria such as BIC [10].

## 3 Bayesian non-parametric parsimonious clustering

Bayesian non-parametric (BNP) mixture approaches for clustering offer a principled alternative to tackle this problem by inferring the number of clusters from the data in a single run, rather than in a two-stage scheme as in standard model-based clustering [4, 3]. They assume that the observed data are governed by an infinite number of clusters, but only a finite number of them do actually generates the data. This is achieved by assuming a general process as prior on the infinite possible partitions, which is not restrictive as in classical Bayesian inference, in such a way that only a (small) finite number of clusters will be actually active. Such a prior can be the CRP [4]. Several Bayesian non-parametric models have considered the general GMM, that is the infinite Gaussian mixture [3] and the Chinese Restaurant Process (CRP) mixture [4]. In the proposed BNP parsimonious clustering approach, we exploit the eigenvalue decomposition of the cluster covariance matrices as in [5, 6] and integrate it into an infinite mixture modeling framework. This leads to an infinite parsimonious Gaussian mixture (IPGMM) which is very flexible in terms of modeling, and automatically infers the number of flexible clusters from the data. We assume a CRP prior over the infinite possible partitions.

### 3.1 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 joint distribution of the unknown cluster labels: $p(\mathbf{z}) = p(z_1)p(z_2|z_1)\ldots p(z_n|z_1, z_2, \ldots, z_{n-1})$. 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 these tables. Customers are social, so that the $i$th 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 $\alpha$ which represents the CRP concentration parameter. This can be explicitly formulated as follows

$$p(z_i = k|z_1, ..., z_{i-1}) = \mathrm{CRP}(z_1, \ldots, z_{i-1}; \alpha) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_+ \end{cases} \tag{3}$$

where $K_+$ is the number of tables for which the number of customers sitting in is $n_k > 0$, and $k \leq K_+$ means that $k$ is a previously occupied table and $k > K_+$ means $k$ is 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 $\mathrm{CRP}(z_1, \ldots, z_{i-1}; \alpha)$ is completed with a likelihood with parameters $\boldsymbol{\theta}$ (i.e., in the GMM case a multivariate Gaussian likelihood), with each table (cluster), and a prior distribution $(G_0)$ for the parameters. For example in the GMM case one can use conjugate priors, that is a multivariate normal inverse-Wishart prior distribution for the mean vectors and the covariance matrices. This means that, the $i$th customer, after sitting at table $z_i = k$, chooses a dish (the parameter $\boldsymbol{\theta}_{z_i}$) from the prior of that table (cluster). This can be summarized by the following generative process.

$$\begin{aligned} z_i &\sim \mathrm{CRP}(z_1, \ldots, z_{i-1}; \alpha) & (4) \\ \boldsymbol{\theta}_{z_i} &\sim G_0 & (5) \\ \mathbf{x}_i &\sim p(.|\boldsymbol{\theta}_{z_i}) \cdot & (6) \end{aligned}$$

According to this process, the generated parameters $\boldsymbol{\theta}_i$ exhibit a clustering property, that is, they share repeated values with positive probability where the unique values of $\boldsymbol{\theta}_i$ shared among the variables are independent draws for the base distribution $G_0$ [4]. The structure of the shared values defines a partition of the integers from 1 to $n$, and the distribution of this partition is a CRP [4].In our proposed infinite parsimonious Gaussian mixture, the parameters $\boldsymbol{\theta}_i$ which include the mean vector and the covariance matrix, the latter is 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.

### 3.2 MCMC Gibbs sampling for model learning

We use a MCMC Gibbs sampling [3, 11, 4] to learn the proposed Bayesian non-parametric parsimonious mixture model. The used priors on the model parameters depends on the type of the parsimonious model. Thus, sampling the model parameters varies according to the considered parsimonious mixture model. Indeed, 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 1 summarizes the considered parsimonious models and the corresponding prior for each model used in the Gibbs sampling.

| Nr. | Decomposition | Model-Type | Prior | Applied to |
|---|---|---|---|---|
| 1 | $\lambda\mathbf{DAD}^T$ | General | $\mathcal{IW}$ | $\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$ |
| 2 | $\lambda_k\mathbf{DAD}^T$ | General | $\mathcal{IG}$ and $\mathcal{IW}$ | $\lambda_k$ and $\mathbf{\Sigma} = \mathbf{DAD}^T$ |
| 3 | $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ | General | $\mathcal{IW}$ | $\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ |
| 4 | $\lambda\mathbf{A}$ | Diagonal | $\mathcal{IG}$ | each diagonal element of $\lambda\mathbf{A}$ |
| 5 | $\lambda_k\mathbf{A}$ | Diagonal | $\mathcal{IG}$ | each diagonal element of $\lambda_k\mathbf{A}$ |
| 6 | $\lambda\mathbf{I}$ | Spherical | $\mathcal{IG}$ | $\lambda$ |
| 7 | $\lambda_k\mathbf{I}$ | Spherical | $\mathcal{IG}$ | $\lambda_k$ |

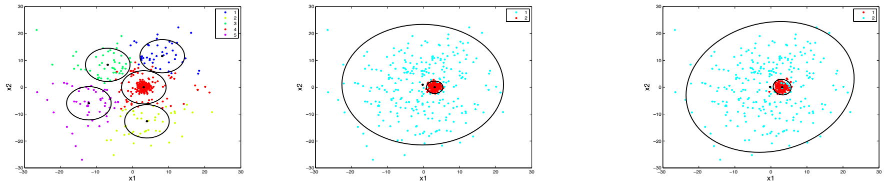
Table 1: Infinite Parsimonious GMMs (IPGMM) with eigenvalue decomposition and the associated prior for each decomposition. Note that $\mathcal{I}$ denotes an inverse distribution, $\mathcal{G}$ a Gamma distribution and $\mathcal{W}$ a Wishart distribution.

## 4 Experiments

We performed experiments on both simulated and real data in order to assess the behavior of our proposed non-parametric method. We highlight its flexibility in terms of modeling, and its use for clustering and selecting the number of clusters. In the experiments, each Gibbs is run ten times with different initializations, each Gibbs run generates 2000 samples. The best solution corresponding to the highest posterior probability is then selected.

## 4.1 Experiment on simulated data

We considered a two-class situation which is the same as for the parametric approach in [6], and 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 $\boldsymbol{\mu}_2 = (3, 0)^T$, $\boldsymbol{\Sigma}_1 = 100\ \mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \mathbf{I}_2$. Figure 1 shows the simulated data and the obtained partitions by the proposed Bayesian non-parametric clustering approach for three different parsimonious models. First, it can be observed that, the partition provided by the spherical model ($\lambda\mathbf{I}$) which does not allow clusters with different volumes, is far from the actual partition. This model also fails for the finite GMM case [6]. 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 error rate equals 4.40%) 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 error rate equals 4.80%). Furthermore, for this simulated data, the best log-likelihood value 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.



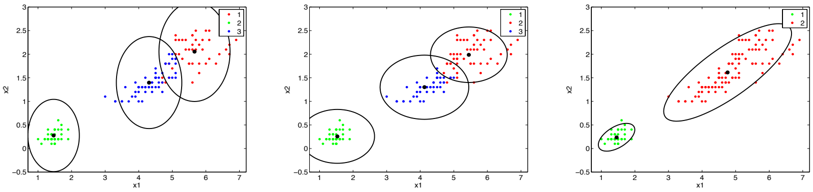
$\lambda\mathbf{I}$: $\hat{K} = 5, \hat{\mathcal{L}} = -57159$ $\lambda_k\mathbf{I}$: $\hat{K} = 2, \hat{\mathcal{L}} = \mathbf{-51125}$ $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$: $\hat{K} = 2, \hat{\mathcal{L}} = -51152$

Figure 1: A two-class data set and the log-likelihood values ($\hat{\mathcal{L}}$), and estimated number of clusters ($\hat{K}$) obtained by infinite parsimonious GMMs.

## 4.2 Experiment on real data

We considered the well-known Iris data set for illustration. Let us recall that Iris data contains 150 data of dimension 4 covering three classes. Figure 2 shows the partition and densities estimated by the proposed non-parametric parsimonious clutering approach.

We can see that both the spherical model $\lambda_k\mathbf{I}$ and the diagonal model $\lambda_k\mathbf{B}$ provide the correct number of classes and allow to reconstruct the hidden data structure. The misclassification error rate for the diagonal model is 5.33% and the one for the spherical model is 10.66%. Let us also note that, for the finite GMM clustering approach, the models which provide the correct number of clusters are the diagonal models $\lambda\mathbf{B}$ and $\lambda_k\mathbf{B}$ and the corresponding misclassification error rates are respectively 9.33% and 11.33%. This can make more advantageous this non-parametric alternative.



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

Figure 2: Clustering results obtained by a spherical model (left), a diagonal model (middle) and the general model (right).

## 5 Conclusion

In this paper we presented a new Bayesian non-parametric parsimonious approach for clustering. It is based on an infinite GMM with a CRP prior and an eigenvalue decomposition of the cluster covariance matrix. It allows deriving several flexible models and avoids the problem of model selection encountered in maximum likelihood and Bayesian learning of parametric GMM. The obtained results highlight the interest of using this infinite parsimonious Bayesian clustering as a good alternative to finite Gaussian clustering. Our current work investigate additional experiments on both simulated and real data and future work may concern other MCMC techniques to learn the models.

## References

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

[2] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *JASA*, 97:611–631, 2002.

[3] C. Rasmussen. The infinite gaussian mixture model. *Advances in neuronal Information Processing Systems*, 10:554 – 560, 2000.

[4] J. Gershman Samuel and David M. Blei. A tutorial on bayesian non-parametric model. *Journal of Mathematical Psychology*, 56:1–12, 2012.

[5] Jeffrey D. Banfield and Adrian E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803–821, 1993.

[6] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793, 1995.

[7] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley, 1997.

[8] C. Fraley and A. E. Raftery. Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of Classification*, 24(2):155–181, 2007.

[9] H. Bensmail, G. Celeux, A. E. Raftery, and C. P. Robert. Inference in model-based cluster analysis. *Statistics and Computing*, 7(1):1–10, 1997.

[10] G. Schwarz. Estimating the dimension of a model. *Ann. of Stat.*, 6:461–464, 1978.

[11] F. Wood, Thomas L. Griffiths, and Z. Ghahramani. A non-parametric bayesian method for inferring hidden causes. In *UAI*, 2006.