Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture

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

[1]Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
[2]Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
[3]Institut Universitaire de France, France

Abstract

In this work we propose to extend the finite parsimonious Gaussian mixture to the infinite case so that the classification of our data could be performed in one stage. We implemented the eigenvalue decomposition of the covariance matrix of each cluster to the Infinite Gaussian mixture model and made it parsimonious. We developed an MCMC algorithm (Gibbs sampling) to learn the various models and we named this approach the bayesian non-parametric parsimonious approach for cluster analysis. The new approach will be more flexible in terms of modeling and will automatically provide the partition of the data and the number of clusters. This approach will be applied into the challenging problem of Whale song decomposition NIPS4B challenge. These algorithms would also give efficient clustering on complex sequence of pulses, and then may allow muti-source/multi-animals labelling.

1 Introduction

Clustering is one of the essential tasks in machine learning and statistics. One of the main problem in data analysis is to estimate the number of clusters that fits best the data. For that we find different approaches in the literature, where one of the most popular is the model-based clustering [1, 2]. These finite parsimonious Gaussian mixtures rely on the eigenvalue decomposition of the covariance matrix, allowing the models to change between the simplest spherical one to the more general [3]. The model parameters can be estimated in a Maximum Likelihood (ML) framework by the Expectation Maximization (EM) algorithm [4] or in a Maximum A Posteriori estimation (MAP) [5] framework or by using MCMC sampling techniques[6, 7]. In this approach, as well as in standard model-based clustering techniques, the selection of the number of clusters is performed by using penalized likelihood criteria such as the Bayesian Information Criteria (BIC) [8], Akaike Information Criterion [9], Integrated Classification Likelihood (ICL)[10], etc. So we need to perform a two stages for classification, first estimate the number of clusters and then run th EM algorithm for the classification of the data.

An alternative well-principled approach for the difficult problem of model selection is to use the Bayesian Non-Parametric (BNP) [11] methods for clustering, one of them being the infinite Gaussian mixture model (IGMM) [12]. Indeed, the principle of IGMM is based on the one of the Chinese Restaurant Process (CRP) [13, 14, 15, 16, 17] which is well suited to the problem of non-parametric clustering. This alternative gives us the possibility to obtain the number of clusters in the same stage of clustering so that as the new data will be observed the number of model parameters can be changed. The general (full GMM) model used in IGMM is not so flexible as in the case of the model-based clustering [3, 18] where the covariance matrix can take different forms, depending on the volume shape and orientation. Therefore we proposed to develop a new approach that will rest being an infinite Gaussian mixture model approach that will give us the possibility to automatically provide the number of classes but know with an eigenvalue decomposition of the covariance matrix giving more flexibility for the model.

The paper is organized as follows. Section 2 briefly discusses previous work on finite Gaussian mixture clustering, in particular we show the model-based clustering approach. Then, Section 3, presents the proposed approach and Section 4 shows experiment results after application to the Whale song decomposition NIPS4B challenge of the EM algorithm with ML and MAP frameworks and the proposed bayesian non-parametric parsimonious approach.

2 Parametric parsimonious Gaussian clustering

It is supposed that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a sample of n i.i.d observations in $\mathbb{R}^d$, and $\mathbf{z} = (z_1, \ldots, z_n)$ is 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.

In the model-based clustering [1, 2, 5] the data X is proposed to be generated from a mixture model with the density:

$$p(\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{x}_i; \theta_k) \tag{1}$$

having f_k a distribution with parameters θ_k and the non-negative mixing proportions π_k that sum to one.

We will suppose in particular the multivariate Gaussian Mixture Model (GMM) [1] to cluster the data X so that in this case we have f_k being a multivariate Gaussian distribution (equation 2) with the parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ which are respectively the mean vector and the covariance matrix for the kth Gaussian component density.

$$f_k(\mathbf{x}_i|\theta_k) = \mathcal{N}_k(\mathbf{x}_i|\mu_k, \Sigma_k) \equiv 2\pi|\Sigma_k|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)\right\} \tag{2}$$

The finite parsimonious GMM by the eigenvalue decomposition of the covariance matrix makes the model more flexible, giving a possibility to variate each cluster density by volume, orientation and shape. The parametrization of the covariance matrix is given in equation 3.

$$\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{3}$$

where λ_k is a scalar that defines the volume, $\mathbf{D}_k$ a orthogonal matrix that defines the orientation and $\mathbf{A}_k$ is a diagonal matrix with determinant 1 witch defines the shape. This decomposition leads to fourteen flexible models [3] going from simplest spherical models to the complex general one.

One of the most used algorithm for learning the model is the Expectation Maximization(EM) algorithm that maximizes the likelihood [19, 20] is an iterative algorithm consisting of two stages, the expectation of the complete data log-likelihood named the E-step and the maximization of the expected complete data log-likelihood named the M-step. Maximizing the likelihood (ML framework) will maximize the mixture likelihood $p(\mathbf{X}|\pi_k, \boldsymbol{\mu}_k, \Sigma_k)$.

$$p(\mathbf{X}|\pi_k, \boldsymbol{\mu}_k, \Sigma_k) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)$$

The maximizing of the posteriori (MAP framework) can be also performed by the EM algorithm [5]. It leads by adding a prior to the mixtures parameters so that it maximizes the following posterior parameter distribution $p(\boldsymbol{\theta}|\mathbf{X})$

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

where $p(\boldsymbol{\theta}) = p(\Sigma)p(\boldsymbol{\mu})$ is the prior for the parameters of the mixture. Also we find in the literature different extension of the EM algorithm like CEM, GEM, etc. that could also be used to learn the model. Another alternative to learn the models are the Markov Chain Monte Carlo (MCMC) algorithms (like Gibbs sampling) [7, 21, 22].

However before learning the model with one of these finite gaussian mixture model we must have the answer to what is the number of mixtures in our model. For that we pose K_{max} that is a maximum number of cluster possible and we compute the penalized log-likelihood criteria (BIC, AIC, ICL, etc.) After choosing the optimal number of clusters that fit best the data we can run one of the learning algorithms.

3 Bayesian non-parametric parsimonious clustering

First off all we make attention that the term of non-parametric learning does't mean at all that the model doesn't have parameters, indeed it means that it could have an infinite number of them as the data grows, in other words it is assumed that the observed data are governed by an infinite number of clusters, but only a finite number of them does actually generates the data. Bayesian non-parametric (BNP) mixtures for clustering offers a good alternative to infer the number of clusters form data within one stage, rather then in two stages like in the case of the parametric modeling [11, 23, 24, 12]. BNP approach proposes to pose a prior on an infinite partitions in such a way that a finite number of clusters will be active. We could use the Chinese Restaurant Process (CRP) prior [25, 26, 23] or a Dirichlet Process Mixture (DPM) [27, 23, 28].

In this work we proposed to develop the previous work called the infinite Gaussian mixture model [12], based on the full GMM, by extending it to a more flexible mixture model where the covariance matrix has an eigenvalue decomposition [3, 18]. We call the new approach the bayesian non-parametric parsimonious approach. We assumed the Chinese Restaurant Process (CRP) prior for the cluster assignments.

Indeed CRP provides a distribution on the infinite partitions of the data, that is a distribution over the positive integers $1, \ldots, n$. Considering the following joint distribution of the unknown cluster assignments: $p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2)\ldots p(z_n|z_1, z_2, \ldots, z_{n-1})$ we can compute each term by using the CRP distribution. The problem of the Chinese Restaurant Process can be expressed by a real human situation if supposing a restaurant that could be extended in a real time by having the possibility to add an infinite number of tables if the number of customers grows. So the CRP is explained as follows: supposing we have this kind of restaurant where one customer is visiting it. This customer enters and sits at the first table. When the second customers enters the restaurant he will sit with a probability $\frac{1}{1+\alpha}$ to the first table and with probability $\frac{\alpha}{1+\alpha}$ to the second table where α will be a dispersion parameter. Going future we say that the n-th customer will be sitting at a new table with a probability equal to $\frac{\alpha}{n-1+\alpha}$ or at the table k with the probability $\frac{n_k}{n-1+\alpha}$ where n_k is the number of customers sitting at table k. The idea of this model is that humans adaptively learn the number of categories of their observations. In the clustering problem the customers are the the observations, so that the new observations can enter the clustering method and choices the table meaning the cluster. This can be explicitly formulated as follows

$$p(z_i = k|z_1, ..., z_{i-1}) = \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) = \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_+ \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_+ \end{cases} \tag{4}$$

where K_+ is the number of tables that have customers sitting on that table $n_k > 0$ or it is also known as active classes. We note $k \leq K_+$ when the k-th table is occupied or in clustering problem the new data observed will be associated to the k-th cluster and $k > K_+$ when a new table will be occupied or the new observation will form a new cluster.

It is also used a prior for the mixture parameters as in MAP approach or the MCMC Gibbs sampling. This priors are used to be conjugate priors so that for example we have the normal inverse-Wishart prior distribution for the mean and the covariance matrix if we use a full GMM. We note this prior distribution as G_0 so that we can show the following generative process.

$$\boldsymbol{\theta}_i \sim G_0 \tag{5}$$

$$z_i \sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \tag{6}$$

$$\mathbf{x}_i \sim p(.|\boldsymbol{\theta}_{z_i}) \cdot \tag{7}$$

According to this generative process we see that $\boldsymbol{\theta}_i$ exhibit a clustering property so that the unique values of the parameters are the number of mixtures that fits the data. G_0 is called the base distribution [27, 23]. The distribution over the partition z_i as it was talked before is a CRP distribution. We proposed to develop the infinite parsimonious Gaussian mixture, where the covariance matrix is parameterized in term of eigenvalue decomposition to provide more flexibility of this model. So the priors on the parameters depends on the type of the parsimonious model. Having chosen the MCMC Gibbs sampling [12, 29, 16, 23] for learning the model we will have different sampling depending on the covariance matrix decomposition.

Indeed, yet we investigated seven parsimonious models, covering the three families of the mixture models which are the general, the diagonal and the spherical family. The parsimonious models therefore go from the simplest spherical one to the more general full model. In table 1, we summarize the considered models and the corresponding prior for each model used in Gibbs sampling.

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

Table 1: Considered Parsimonious GMMs via eigenvalue decomposition and the associated prior distribution for the covariance. Note that $\mathcal{I}$ means that it is an inverse distribution, $\mathcal{G}$ means that it is a Gamma distribution and $\mathcal{W}$ means that it is a Wishart distribution.

4 Experiments

We compared our Bayesian non-parametric parsimonious mixture with model-based clustering (ML-based and MAP-based) approaches. For the ML and MAP approaches, we used the EM algorithm to estimate the model parameters. The model selection is performed by ICL for values of K between 1 and 60. For each value of K, we considered 10 runs of EM, with different initializations, to estimate the mixture model parameters and the one providing the best solution (corresponding to the maximum value of the log-likelihood is selected). Then, the value of K corresponding to the highest ICL value is considered as the best solution with the optimal number of clusters.

For the Bayesian non-parametric approach (IGMM), we used the Gibbs sampler by running it ten times and selecting the best solution in the sense of the posterior.

We illustrate the estimations of the number of classes for Gibbs samplings for 2 spherical models $\lambda\mathbf{I}$ and $\lambda_k\mathbf{I}$, 2 diagonal models $\lambda\mathbf{B}$ and $\lambda_k\mathbf{B}$ and two general models $\lambda_k\mathbf{DAD}^T$ and $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ in the histograms of figure 1. Note that we dont take in consideration the first 50 iterations of the Gibbs sampling. For this whale song data we can conclude that for these models we have been estimated a different number of clusters, that could be compared when estimating the number of clusters by using the information criteria.

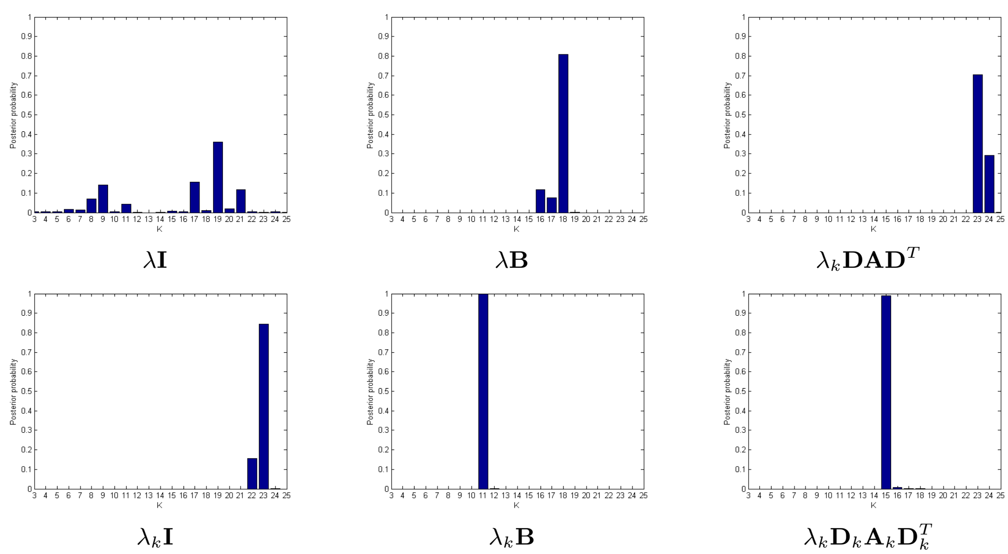


Figure 1: Posterior distribution of the number of clusters obtained by the proposed bayesian non-parametric approach.

The table 2 shows the log-likelihood values that are divided by 10^6 and the number of estimated classes obtained by using the Expectation Maximization (EM) algorithm with one of the information criteria and the proposed bayesian non-parametric parsimonious method for clustering the data. By analysing the results we can conclude that the best solution is by using the more general model with the eigenvalues decomposition of the covariance matrix $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$, meaning that the volume, the orientation and the shape can vary for each cluster. The best likelihood obtained here is by using the EM with maximum a posteriori framework algorithm, that estimates 18 classes. On the other hand, the bayesian non-parametric model estimates 15 classes. By using the spherical models, the one with the equal volumes $\lambda\mathbf{I}$ and the other one with different volumes $\lambda_k\mathbf{I}$, we notice that the estimation of classes are taken to be the maximum, equal to 60, when using the finite Gaussian Mixture Models (GMM), while for the infinite case we have estimated 9 classes for the $\lambda\mathbf{I}$ model and 23 classes for the $\lambda_k\mathbf{I}$ model. Also, for the diagonal models, we have the model with equal volumes $\lambda\mathbf{B}$ that estimates 22 classes for the finite mixture models when using the EM ML approach or EM MAP approach with the Integrated Classification Likelihood (ICL) criteria, and 18 classes when using the proposed non-parametric bayesian clustering. [1]

Table 2: Log-likelihood values (divided by 10^6) and the number of estimated classes obtained for the whale song data set by using the Expectation Maximization approach with maximization of the likelihood (ML) approach and with the maximization a posteriori (MAP) approach and the proposed bayesian parsimonious approach (IPGMM).

	EM ML		EM MAP		IPGMM	
Model	$\hat{K}$	log-lik	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda\mathbf{I}$	60	-2.2198	60	-2.1924	9	-2.3413
$\lambda_k\mathbf{I}$	60	-2.1129	60	-2.0858	23	-2.2133
$\lambda\mathbf{B}$	22	-2.1435	22	-2.1339	18	-2.1958
$\lambda_k\mathbf{B}$	59	-2.0059	53	-1.9595	11	-2.1900
$\lambda\mathbf{DAD}^T$	-	-	34	-2.0815	33	-2.1695
$\lambda_k\mathbf{DAD}^T$	51	-1.9811	-	-	24	-2.1589
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	19	-1.9418	18	$\mathbf{-1.9381}$	15	-2.1234

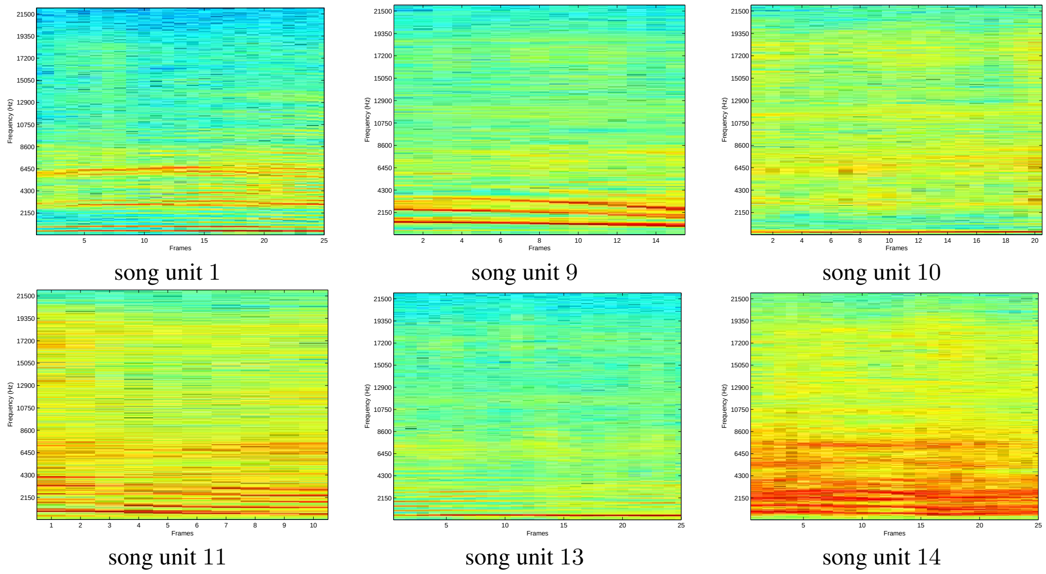
In the figure 2 we show the spectrograms of the whale songs obtained with the proposed bayesian non-parametric approach with the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$. We chose to show these spectrograms of the whale songs because we obtained the best log-likelihood solution when using the new method. On the vertical axes the frequency is showed and on the horizontal axes we have the frames, each frame being represented by 10 ms. As we observe in the table 2 we have 15 clusters for the $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ model when using the infinite Gaussian mixture model, so in figure 2 we show the 6 spectrograms of the whale songs that the time repass 10 ms.

By classification the whale song data with the infinite gaussian mixture model using the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ we see in the figure 3 the song that where observed for each observation. The songs (classes) 8, 12 and 15 are uniformly activated in time, therefore we may figure out that they are representing the sea noise. Whereas the songs (classes) 10,13 and 14 are clearly conveying information (low entropy).

5 Conclusion

This work presents a new Bayesian non-parametric 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 in maximum likelihood-based and Bayesian parametric Gaussian mixture. We applied this method on the Whale song decomposition NIPS4B challenge. The obtaining results highlight the interest of using parsimonious Bayesian clustering as a good alternative namely to finite parsimonious GMM clustering. We saw that the infinite parsimonious Gaussian mixture model (IPGMM) is

[1] The missing values for the two state of art models ($\lambda\mathbf{DAD}^T$ model for EM ML and the $\lambda_k\mathbf{DAD}^T$ model for EM MAP) are due to some trobles when executing the em algorithm for this data and are currently being fixed.



song unit 1 song unit 9 song unit 10

song unit 11 song unit 13 song unit 14

Figure 2: Spectrograms for the whale songs obtained with the proposed bayesian non-parametric approach with the most general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$.

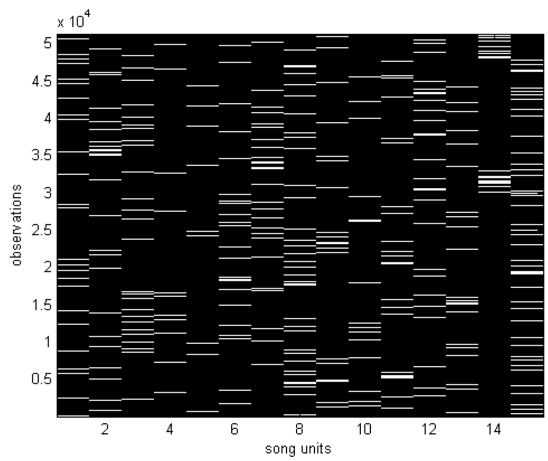


Figure 3: Clusters activities versus time sea noise obtained by IPGMM with $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ model

more flexible in terms of modeling and automatically provides a partition of the data and the number of clusters for the data needed to be clustered.

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*, 97:611–631, 2002.

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

[4] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Comput. Stat. Data Anal.*, 14(3):315–332, October 1992.

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

[6] Halima Bensmail. *Regularized models in discrimination and Bayesian classification*. PhD thesis, University Paris 6, 1995.

[7] 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.

[8] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[9] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[10] C. Biernacki, G. Celeux, and G Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2000.

[11] N. Hjort, Holmes C., P. Muller, and S. G. Waller. *Bayesian Non Parametrics*. Cambrige University Press, 2010.

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

[13] E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Phd. thesis, MIT, Cambridge, MA, 2009.

[14] D. Görür. *Nonparametric Bayesian Discrete Latent Variable Methods for Unsupervised Learning*. Phd. thesis, Berlin Institute of Technology, Berlin, D83, 2007.

[15] Xiaodong Yu. Gibbs sampling methods for dirichlet process mixture model: Technical details. Technical report, University of Maryland, College Park, University of Maryland, College Park, September 2009.

[16] F. Wood, Thomas L. Griffiths, and Z. Ghahramani. A non-parametric bayesian method for inferring hidden causes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2006.

[17] F. Wood and M. J. Black. A nonparametric bayesian alternative to spike sorting. *Journal of neuroscience methods*, 173(1):1–12, 2008.

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

[19] 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*, 39(1):1–38, 1977.

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

[21] H. Bensmail and J. J. Meulman. Model-based clustering with noise: Bayesian inference and estimation. *J. Classification*, 20(1):049–076, 2003.

[22] D. Ormoneit and V. Tresp. Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9(4):639–650, 1998.

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

[24] Erik B. Sudderth. *Graphical models for visual object recognition and tracking*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 2006.

[25] D.J. Aldous. Exchangeability and related topics. In *École d'Été St Flour 1983*, pages 1–198. Springer-Verlag, 1985. Lecture Notes in Math. 1117.

[26] J. Pitman. Combinatorial stochastic processes. Technical Report 621, Dept. of Statistics. UC, Berkeley, 2002.

[27] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

[28] C. E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

[29] R. M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993.