Bayesian non-parametric parsimonious mixtures for model-based clustering

Marius Bartcus

Université de Toulon

Ecole doctorale 548

UMR CNRS LSIS - DYNI team

THÈSE

présentée en vue de l'obtenition du Grade de

Docteur de l'Université de Toulon

Spécialité: Informatique et Mathématiques Appliquées

par

Marius BARTCUS

Bayesian non-parametric parsimonious mixtures for model-based clustering

---

Soutenue publiquement le 26 octobre 2015 devant le jury composé de :

M. Younès BENNANI	Professeur, Université Paris 13	(Rapporteur)
M. Christophe BIERNACKI	Professeur, Université Lille 1, INRIA	(Rapporteur)
M. Allou SAMÉ	Chargé de recherche HDR, IFSTTAR	(Examinateur)
M. Badih GHATTAS	Maître de Conférences HDR, Aix Marseille Université	(Examinateur)
M. Hervé GLOTIN	Professeur, Université de Toulon	(Directeur)
M. Faicel CHAMROUKHI	Maître de Conférences, Université de Toulon	(Encadrant)

Acknowledgments

First, I would like to express my greatest thanks to my advisor, M.Faicel CHAMROUKHI, who guided and inspired me. His support, availability and pacience all along these years contributed a lot in writing my dissertation.

Special thanks to my director, M. Hervé GLOTIN, for his guidance.

I would also like to express my gratitude to M. Younès BENNANI and M. Christophe BIERNACKI for accepting to review my thesis and for their valuable examination.

I am greatly thankful to M. Allou SAMÉ and M. Badih GHATTAS, that accepted to be part of my committee.

Finally, I express my thanks to my family and friends, especially to my wife Diana and my mother Margarita. Without their care, love, moral support, I surely could not complete my doctoral degree.

Marius BARTCUS
Université de Toulon
La Garde, 20 october 2015

To my family, A ma famille

Résumé

Cette thèse porte sur l'apprentissage statistique et l'analyse de données multi-dimensionnelles. Elle se focalise particulièrement sur l'apprentissage non supervisé de modèles génératifs pour la classification automatique. Nous étudions les modèles de mélanges Gaussians, aussi bien dans le contexte d'estimation par maximum de vraisemblance via l'algorithme EM, que dans le contexte Baéyesien d'estimation par Maximum A Posteriori via des techniques d'échantillonnage par Monte Carlo. Nous considérons principalement les modèles de mélange parcimonieux qui reposent sur une décomposition spectrale de la matrice de covariance et qui offre un cadre flexible notamment pour les problèmes de classification en grande dimension. Ensuite, nous investigons les mélanges Bayésiens non-paramétriques qui se basent sur des processus généraux flexibles comme le processus de Dirichlet et le Processus du Restaurant Chinois. Cette formulation non-paramétrique des modèles est pertinente aussi bien pour l'apprentissage du modèle, que pour la question difficile du choix de modèle. Nous proposons de nouveaux modèles de mélanges Bayésiens non-paramétriques parcimonieux et dérivons une technique d'échantillonnage par Monte Carlo dans laquelle le modèle de mélange et son nombre de composantes sont appris simultanément à partir des données. La sélection de la structure du modèle est effectuée en utilisant le facteur de Bayes. Ces modèles, par leur formulation non-paramétrique et parcimonieuse, sont utiles pour les problèmes d'analyse de masses de données lorsque le nombre de classe est indéterminé et augmente avec les données, et lorsque la dimension est grande. Les modèles proposés validés sur des données simulées et des jeux de données réelles standard. Ensuite, ils sont appliqués sur un problème réel difficile de structuration automatique de données bioacoustiques complexes issues de signaux de chant de baleine. Enfin, nous ouvrons des perspectives Markoviennes via les processus de Dirichlet hiérarchiques pour les modèles Markov cachés.

Mots-clés: Apprentissage non-supervisé, modèles de mélange, classification automatique, mélanges parcimonieux, modèles de mélanges bayésiens non-paramétriques, processus de Dirichlet, sélection Bayésienne de modèle

Abstract

This thesis focuses on statistical learning and multi-dimensional data analysis. It particularly focuses on unsupervised learning of generative models for model-based clustering. We study the Gaussians mixture models, in the context of maximum likelihood estimation via the EM algorithm, as well as in the Bayesian estimation context by maximum a posteriori via Markov Chain Monte Carlo (MCMC) sampling techniques. We mainly consider the parsimonious mixture models which are based on a spectral decomposition of the covariance matrix and provide a flexible framework particularly for the analysis of high-dimensional data. Then, we investigate non-parametric Bayesian mixtures which are based on general flexible processes such as the Dirichlet process and the Chinese Restaurant Process. This non-parametric model formulation is relevant for both learning the model, as well for dealing with the issue of model selection. We propose new Bayesian non-parametric parsimonious mixtures and derive a MCMC sampling technique where the mixture model and the number of mixture components are simultaneously learned from the data. The selection of the model structure is performed by using Bayes Factors. These models, by their non-parametric and sparse formulation, are useful for the analysis of large data sets when the number of classes is undetermined and increases with the data, and when the dimension is high. The models are validated on simulated data and standard real data sets. Then, they are applied to a real difficult problem of automatic structuring of complex bioacoustic data issued from whale song signals. Finally, we open Markovian perspectives via hierarchical Dirichlet processes hidden Markov models.

Keywords: Unsupervised learning, mixture models, model-based clustering, parsimonious mixtures, Dirichlet process mixtures, Bayesian non-parametric learning, Bayesian model selection

Contents

Notations

For more understanding we shall list the notations that are used in this thesis. A vector will be written in bold. (e.g, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}\dots$). I will assume that all vectors are column vectors, so that the transpose of a column vector $\mathbf{x}$, noting $\mathbf{x}^T$ is a row vector. Matrices are also notated in a bold manner (e.g, $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}\dots$). The transpose of a matrix $\mathbf{X}$ is notated as $\mathbf{X}^T$. Future we shall suppose that a matrix have n rows and d columns. An identity matrix with size n is noted by $\mathbf{I}$.

General Notations

$L(\mathbf{X}|\boldsymbol{\theta})$ the likelihood of the function of parameter vector $\boldsymbol{\theta}$ for the data $\mathbf{X}$

$L_c(\mathbf{X}|\boldsymbol{\theta})$ the complete likelihood of the function of parameter vector $\boldsymbol{\theta}$ for the data $\mathbf{X}$

$\mathrm{tr}(\mathbf{A})$ trace of $\mathbf{A}$

$\mathrm{diag}(\mathbf{A})$ diagonal terms of matrix $\mathbf{A}$

Multidimensional Data

$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ a sample with n observations, each sample having d features.

$\mathbf{x}_i$ ith observation

$\mathbf{z} = (z_1, \dots, z_n)$ hidden class vector

K number of components (clusters)

$z_i = k \in \{1, \dots, K\}$ class label for $\mathbf{x}_i$

Probability distribution

$p(.)$ generic notation of a probability density function (p.d.f)

$\mathcal{I}$ an inverse distribution

$\mathcal{N}$ Gaussian (normal) distribution

$\mathcal{W}$ Wishart distribution

$\mathcal{G}$ Gamma distribution

Mult(.) Multinomial distribution

Dir(.) Dirichlet distribution

Graphical model representation Figure 1 gives the convention for the probabilistic graphical models in this thesis. The gray circles will denote observed continuous variables, the dots denote deterministic variables and the circles will denote observed continuous variables. The arrows describe the conditional dependence between variables. Finally, the rectangle denotes the variable repetitions, with the specified number of repetitions.

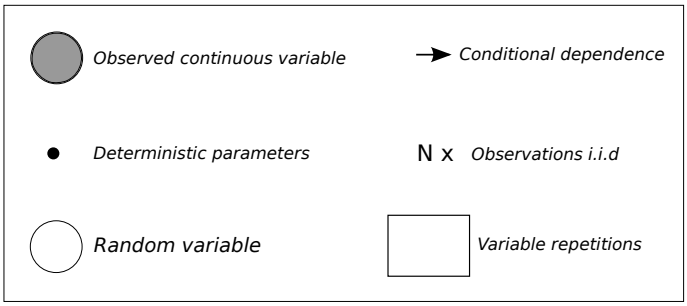


Figure 1: *Graphical model representation conventions.*

- Chapter 1 -

Introduction

Le travail présenté dans cette thèse s'inscrit dans le cadre général de l'apprentissage statistique (Mitchell, 1997; Vapnik, 1999; Vapnik and Chervonenkis, 1974) à partir de données complexes. En particulier, nous nous sommes intéressés à l'apprentissage de modèles génératifs (Jebara, 2001, 2003) pour l'analyse de données multidimensionnelles dans un contexte non-supervisé. Dans ce contexte, les observations sont souvent incomplètes et il y a donc nécessité de reconstruire l'information manquante. C'est le cas en classification automatique qui est au coeur de cette thèse. En apprentissage génératif non-supervisé, les modèles à variables latentes, en particulier les modèles de mélange (Frühwirth-Schnatter, 2006; McLachlan and Basford, 1988; McLachlan and Peel., 2000; Titterington et al., 1985) ou leur extension pour les données séquentielles, tel que les modèles de Markov cachés (Frühwirth-Schnatter, 2006; Rabiner, 1989), fournissent un cadre statistique pertinent pour une telle analyse de données incomplètes. Nous nous sommes focalisé sur le problème de modélisation de données hétérogènes, se présentant sous forme de sous-populations, à travers des modèles de mélanges de densités. Les modèles de mélange offrent en effet un cadre pertinemment flexible pour la classification automatique "clustering", l'un des principaux sujets d'analyse traité dans cette thèse. Le clustering est un problème largement étudié en statistique et en apprentissage automatique ainsi que dans beaucoup d'autres domaines connexes. Le problème de la classification automatique est abordé ici en utilisant des mélanges (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 1998a; McLachlan and Basford, 1988; Scott and Symons, 1981).

La classification automatique à base de modèles de mélange, en anglais "model-based clustering", consiste en l'estimation de densité et nécessite donc la construction de bon estimateurs. Ce problème d'apprentissage des modèles est étudié aussi bien dans le paradigme fréquentiste en reposant sur l'estimation par maximum de vraisemblance en utilisant l'algorithme

Espérance-Maximization (EM) (e.g voir McLachlan and Krishnan (2008)), que dans le cadre bayésien (e.g voir Stephens (1997)), en se basant sur l'estimation par maximum a posteriori en utilisant les techniques d'échantillonnage par Monte Carlo (MCMC) (Diebolt and Robert, 1994; Marin et al., 2005; Neal, 1993).

Nous avons étudié le problème d'inférence des modèles de mélanges à partir des deux points de vue, mais nous nous sommes concentrés principalement sur le paradigme bayésien. En effet, l'apprentissage des mélanges par maximum de vraisemblance peut avoir quelques instabilités en pratique en raison des singularités ou des dégénérescences lors de l'estimation de paramètres (Fraley and Raftery, 2007a, 2005; Ormoneit and Tresp, 1998; Snoussi and Mohammad-Djafari, 2000, 2005; Stephens, 1997). La régularisation bayésienne offre une bonne alternative, même si elle est également confrontée à des difficultés pratiques, liées principalement à un coût de calcul qui peut être très significatif en particulier à grande échelle. L'estimation bayésienne offre aussi dans son extension non-paramétrique (Hjort et al., 2010; Navarro et al., 2006; Neal, 2000; Orbanz and Teh, 2010; Rasmussen, 2000), un cadre bien établi à d'autres problématiques pour les modèles de mélange, en particulier la sélection et la comparaison des modèles. L'approche non-paramétrique offre en effet une bonne alternative au problème de sélection de modèle en estimant simultanément le modèle et le nombre de ses composantes à partir des données. Ceci est une alternative à ce qui est classiquement utilisé dans les mélanges finis en choix de modèle, à savoir l'utilisation de critères d'information tels que le critère d'information bayésien (BIC) (Schwarz, 1978), le critère d'information d'Akaike Akaike (1974) ou le critère de la vraisemblance classifiante intégrée (ICL) (Biernacki et al., 2000) dans une approche à deux étapes afin de sélection un modèle parmi plusieurs candidats pré-estimés. Dans ce contexte, nous avons étudié l'utilisation de modèles non-paramétriques qui reposent sur des processus généraux flexibles comme a priori, que les processus de Dirichlet (Antoniak, 1974; Ferguson, 1973) ou par équivalence les processus du restaurant chinois (Aldous, 1985; Pitman, 2002; Samuel and Blei, 2012).

D'autre part, il est connu que les mélanges standards, en particulier le mélange Gaussien, comme beaucoup d'autres approches de modélisation, peuvent conduire à des solutions non satisfaisantes, dans le cas de données de grande dimension (Bouveyron, 2006; Bouveyron and Brunet-Saumard, 2014). Le nombre de paramètres à estimer en effet augmente rapidement lorsque la dimension est élevée, ce qui peut rendre l'estimation problématique. Cela a été étudié notamment dans les mélanges parcimonieux qui se basent sur une décomposition spectrale de la matrice de covariance, et qui ont montré leur performance, en particulier classification automatique en analyse fréquentiste (Banfield and Raftery, 1993; Bensmail and Celeux, 1996; Celeux and Govaert, 1995), ainsi qu'en analyse bayésienne paramétrique (Bensmail and Meulman, 2003; Bensmail et al., 1997; Bensmail, 1995; Fra-

ley and Raftery, 2002, 2007a, 2005). Nous avons étudié ces modèles, particulièrement dans le cadre bayésien. Ensuite, nous avons dérivé une approche bayésienne non-paramétrique pour les mélanges parcimonieux ou l'apprentisage du modèle est effectué dans un contexte bayésien non-paramétrique avec des priori flexibles tels que le processus du restaurant chinois, et où le choix du modèle s'effectue par le facteur de Bayes.

Dans le Chapitre 2 dédié à l'état de l'art, nous décrivons les modèles de mélanges pour la classification automatique ainsi que l'estimation des mélanges par maximum de vraisemblance en utilisant l'algorithme EM (Celeux and Govaert, 1995; Dempster et al., 1977; McLachlan and Krishnan, 2008). Nous considérerons le cas général du mélange et nus nous focalisons sur les mélanges Gaussiens, qui sont largement utilisés en analyse statistique. Nous étudions et discutons également des modèles parcimonieux, dérivés du modèle de mélange Gaussien standard. Enfin, nous discutions la problématique classique de la sélection de modèle qui est généralement traitée par des critères de choix sélectionnant un modèle parmi une collection de modèles candidats pré-estimés.

Ensuite, dans le Chapitre 3, nous étudions les mélanges pour la classification automatique dans une contexte bayésien où le but est de traiter les limites de l'approche décrite précédemment. Nous étudions deux approches pour l'apprentissage Bayésien des mélanges. La première consiste à utiliser un algorithme EM bayésien (Fraley and Raftery, 2007a, 2005; Ormoneit and Tresp, 1998; Snoussi and Mohammad-Djafari, 2000, 2005). La seconde consiste quant à elle en la construction d'un estimateur du MAP en utilisant les techniques d'échantillonnage MCMC (Diebolt and Robert, 1994; Geyer, 1991; Gilks et al., 1996; Marin et al., 2005; Neal, 1993; Stephens, 1997). Une attention particulière est portée sur les modèles parcimonieux pour lesquels nous mettons en œuvre plusieurs modèles et effectuons une étude expérimentale comparative pour les évaluer. Aussi, nous étudions le problème de sélection et de comparaison de ces modèles parcimonieux en utilisant des critères d'informations y compris le facteur de Bayes.

Dans le Chapitre 4, nous développons une formulation bayésienne non-paramétrique pour les modèles de mélanges parcimonieux (DPPM). En s'appuyant sur les mélanges de processus de Dirichlet, ou par équivalence les mélanges de processus du restaurant chinois, nous introdusons des modèles parcimonieux de processus de Dirichlet qui fournissent un cadre flexible pour la modélisation de différentes structures des données ainsi qu'une bonne alternative pour résoudre le problème de sélection de modèle. Nous dérivons un échantillonnage de Gibbs pour estimer les modèles et nous utilisons le facteur de Bayes pour la sélection et la comparaison des modèles (Bartcus et al., 2014, 2013; Chamroukhi et al., 2015, 2014b,a).

Ensuite, le Chapitre 5 sera dédié aux expérimentations afin d'évaluer nos modèles. Nous évaluons les modèles bayésiens non-parametriques parcimonieux proposés, ainsi que ceux du cas paramétrique, sur plusieurs jeux

de données simulées et réelles. Une application de traitement non-supervisé de signaux bioacoustiques est aussi étudiée.

Dans le Chapitre 6, nous ouvrons de futures extensions possibles de notre approche DPPM pour l'analyse de séquences. Nous montrons des résultats expérimentaux en appliquant les modèles récents de l'état de l'art de processus Dirichlet hiérarchique pour les mélanges de Markov caché (HDP-HMM) (Beal et al., 2002; Fox, 2009; Fox et al., 2008; Teh and Jordan, 2010; Teh et al., 2006) qui sont bien adaptés aux données séquentielles. Les résultats obtenus mettent en évidence que le cadre bayésien non-paramétrique est bien adapté pour ces données.

Enfin, dans le Chapitre 7 est dédiée à une conclusion et discussions, ainsi que de futures perspectives de recherche possibles liées aux DPPMs.

Introduction

The work presented in this thesis lies in the general framework of statistical learning (Mitchell, 1997; Vapnik, 1999; Vapnik and Chervonenkis, 1974) from complex data, particularly, the generative part of statistical learning (Jebara, 2001, 2003) for multivariate data analysis, that is, to learn from samples of individuals described by vectors in $\mathbb{R}^d$. We are indeed interested in understanding the process generating the data, through the construction of probabilistic models and deriving algorithms for such analysis. We focus on the paradigm in which the analysis is performed in an unsupervised way, that is, in a missing data framework, where the observed individuals are incomplete or require recovering possible hidden information. In such a context, latent data models particularly mixture models (Frühwirth-Schnatter, 2006; McLachlan and Basford, 1988; McLachlan and Peel., 2000; Titterington et al., 1985) or their extensions to sequential data, that is, hidden Markov models (Frühwirth-Schnatter, 2006; Rabiner, 1989) provide a well-established statistical framework for such analysis in an incomplete data context. In particular, we focus on the problem of modeling data which present heterogeneities in the form of several sub-populations. To this end, mixture models, thanks to their flexibility and their sound statistical background, are one of most popular and successful models in this context of analysis. One main topic of analyses, under this mixture modeling context, is cluster analysis, an unsupervised widely studied problem in statistics and machine learning as well as in many other related area. The problem of clustering is tackled here by using mixtures, that is, the so-called mixture model-based clustering framework (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 1998a; McLachlan and Basford, 1988; Scott and Symons, 1981).

In cluster analysis with mixtures, the analysis consists in density estimation, which therefore requires the construction of desirable estimators. This is the problem of fitting mixtures, which is classically addressed from two different, but also related paradigms, that is the frequentist one which relies on the maximum likelihood estimator by using Expectation-Maximization (EM) algorithms (e.g see McLachlan and Krishnan (2008)), and the Bayesian

one (e.g see Stephens (1997)), which provide distributions over the model rather than a point estimation as in the frequentist approach, by relying on the so-called maximum a posteriori (MAP) estimator by using Markov Chain Monte Carlo (MCMC) (Diebolt and Robert, 1994; Marin et al., 2005; Neal, 1993).

We study the problem of fitting mixtures from the two points of view but we mainly focus on the Bayesian paradigm. Indeed, the maximum likelihood fitting of mixtures may be subject to some instabilities in practice due to the singularities or degeneracies of parameter estimates (Fraley and Raftery, 2007a, 2005; Ormoneit and Tresp, 1998; Snoussi and Mohammad-Djafari, 2000, 2005; Stephens, 1997). The Bayesian regularization may offer a good alternative, but also is subject to practical difficulties, mainly related to an important computational load. The Bayesian framework offers, also, under non-parametric extensions (Hjort et al., 2010; Navarro et al., 2006; Neal, 2000; Orbanz and Teh, 2010; Rasmussen, 2000), a well-established framework to other issues in mixture modeling, that is those of model selection and comparison. They offer a well established alternative to the problem of model selection, which is general equivalent to the one of choosing the number of mixture components, by relying on general adapted priors. This is an alternative to the one generally used in finite mixture by using information criteria such as the Bayesian Information Criteria (BIC) (Schwarz, 1978), the Akaike Information Criteria (AIC) Akaike (1974) or the Integrated Classification Likelihood (ICL) (Biernacki et al., 2000) etc. in a two-fold scheme. In this context, we investigate the use of non-parametric models that rely on general flexible priors such as Dirichlet Processes (Antoniak, 1974; Ferguson, 1973) or by equivalence their Chinese Restaurant Process (Aldous, 1985; Pitman, 2002; Samuel and Blei, 2012).

On the other hand, it is known that the standard mixtures, particularly Gaussian mixtures, may lead to non accurate solutions, as many other modeling approaches, in the case of high dimensional data (Bouveyron, 2006; Bouveyron and Brunet-Saumard, 2014). The number of parameters to be estimated may grow up rapidly with the number of components especially when the dimension is high. This was investigated by proposing the parsimonious mixtures by parameterizing the component specific covariance matrix by an eigenvalue decomposition, and which have shown their performance in particular for cluster analysis in the maximum likelihood fitting context (Banfield and Raftery, 1993; Bensmail and Celeux, 1996; Celeux and Govaert, 1995) as well as in parametric Bayesian model-based clustering (Bensmail and Meulman, 2003; Bensmail et al., 1997; Bensmail, 1995; Fraley and Raftery, 2002, 2007a, 2005). We revisit these models from mainly the Bayesian prospective. We investigate the Bayesian parametric case. Then we derive them within a full Bayesian non-parametric approach where both the fitting is tackled in a principled way within a Bayesian formulation by relying on general flexible priors such as Chinese Restaurant Process and

the Dirichlet Process, and the issue of model selection and comparison takes benefit of the well-tailored Bayes Factors.

The outline and the contributions of this thesis are summarized as follows.

In Chapter 2, we provide an account of the state of the art approaches in model-based clustering. We describe the maximum likelihood fitting for mixtures with the Expectation-Maximization (EM) algorithm (Celeux and Govaert, 1995; Dempster et al., 1977; McLachlan and Krishnan, 2008). We consider the general case of mixture and focus on the Gaussian mixture, which is widely used in statistical analysis. We also study the parsimonious models derived from the standard Gaussian mixture model and discuss them. Finally, the classical issue of model selection is discussed in this context where it is in general addressed by external criteria to select a model from a previously fitted collection of model candidates.

Then, in Chapter 3, we investigate the problem of mixture model-based clustering from a Bayesian point of view where the aim is to deal with limitations of the previously described approach. We study the case of Bayesian mixture fitting by examining two ways. The first one consists in using a Bayesian EM (Fraley and Raftery, 2007a, 2005; Ormoneit and Tresp, 1998; Snoussi and Mohammad-Djafari, 2000, 2005), and the second one consists in the construction of a full MAP estimator by using Markov Chain Monte Carlo (MCMC) sampling (Diebolt and Robert, 1994; Geyer, 1991; Gilks et al., 1996; Marin et al., 2005; Neal, 1993; Stephens, 1997). An attention is given to the parsimonious models, for which we implement several models and perform a comparative experimental study to assess them. We also investigate the problem of model selection and comparison of these parsimonious models by using criteria including Bayes Factors (Basu and Chib, 2003; Carlin and Chib, 1995; Gelfand and Dey, 1994; Kass and Raftery, 1995; Raftery, 1996).

In Chapter 4 we develop a Bayesian non-parametric formulation for the parsimonious mixture models. By relying on Dirichlet Process mixtures, or by equivalence the Chinese Restaurant Process mixtures, we introduce Dirichlet Process Parsimonious Mixture (DPPM) models, which provide a flexible framework for modeling different data structures as well as a good alternative to tackle the problem of model selection. We derive a Gibbs sampler to infer the models and use Bayes Factors for model selection and comparison (Bartcus et al., 2014, 2013; Chamroukhi et al., 2015, 2014b,a).

Then Chapter 5 is dedicated for experiments to assess the models. We implemented the presented Bayesian non-parametric parsimonious mixture models, as well as those in the parametric case, and evaluated them on simulated datasets, benchmarks and a real-world data set issued from a bioacoustic signal processing application.

In Chapter 6 in order to open possible future extensions of the proposed Dirichlet Process Parsimonious Mixture models, we show the experimental

results obtained by applying the quiet recent state of the art Hierarchical Dirichlet Process for Hidden Markov Models (HDP-HMM) (Beal et al., 2002; Fox, 2009; Fox et al., 2008; Teh and Jordan, 2010; Teh et al., 2006) which are tailored to sequential data. The obtained results highlight, that the Bayesian non-parametric framework is adapted for such data as it provides encouraging results. Thus, the DPPM which also provide an interesting and encouraging results in such a context of sequential data modeling, are likely to more improve the results if they are extended to the sequential context.

Finally, in Chapter 7 we draw concluding remarks and open possible future research perspectives related to the DPPMs.

- Chapter 2 -

Mixture model-based clustering

Contents

2.1 Introduction

In this chapter we describe state of the art approaches for clustering based on the finite mixture model. The mixture models (Pearson, 1894; Scott and Symons, 1971), in particular the finite mixture models are also named in literature as the parametric model-based clustering (Banfield and Raftery, 1993; Böhning, 1999; Fraley and Raftery, 1998a, 2002; Frühwirth-Schnatter, 2006; Lindsay, 1995; McLachlan and Basford, 1988; McLachlan and Peel., 2000; Titterington et al., 1985).

2.2 The finite mixture model

The finite mixture model is a probabilistic model used in machine learning and statistics to model distributions over observed data organized into groups. It has shown great performance in cluster analysis.

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a sample of n i.i.d observations in $\mathbb{R}^d$. The finite mixture model decomposes the density of the observed data as a weighted sum of a finite number of K component densities. The density function of the data is given by the following mixture density:

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k p_k(\mathbf{x}_i|\boldsymbol{\theta}_k), \tag{2.1}$$

where the π_k's, given by $\pi_k = p(z_i = k)$ are the mixing proportions which represent the probabilities that the data point $\mathbf{x}_i$ belongs to component k. They are non-negative $\pi_k \geq 0, \forall k = 1 \ldots K$ and sum to one, that is $\sum_{k=1}^{K} \pi_k = 1$, $p_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$ is the density function for the kth component with parameters $\boldsymbol{\theta}_k$ and $\boldsymbol{\theta} = \{\pi_1, \ldots \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ are the model mixture parameters

From a generative point of view, the process for generating data from the finite mixture model can be stated as follows. First, a mixture component z_i is sampled independently according to a Multinomial distribution given the mixing proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. Then, given the mixture component $z_i = k$, and the corresponding parameters $\boldsymbol{\theta}_{z_i}$, the data $\mathbf{x}_i$ are generated independently from the supposed distribution $p_k(\mathbf{x}_i|\boldsymbol{\theta}_{z_i})$. The process is repeated n times, with n the number of observations. This generative process for the finite mixture model is summarized by the two steps:

$$\begin{aligned} z_i &\sim \text{Mult}(1; \pi_1, \ldots, \pi_k), \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim p_k(\mathbf{x}_i|\boldsymbol{\theta}_{z_i}). \end{aligned} \tag{2.2}$$

Generally, p_k are distributions from the same family with different parameters. For instance they can all be Poisson distributions (see Rau et al.

(2011)); Gamma distributions (see Almhana et al. (2006); Mayrose et al. (2005)); Bernoulli distributions (see Juan and Vidal (2004); Juan et al. (2004)); Multinomial distributions (see Novovičová and Malík (2003)); Student-t distributions (see McLachlan and Peel. (2000); Peel and McLachlan (2000); Svensen and Bishop (2005); Wang and Hu (2009)); skew normal and skew t-distributions (see Azzalini (1985); Gupta et al. (2004); Lee and McLachlan (2013); Pyne et al. (2009)); the Gaussian (normal) distributions (see Banfield and Raftery (1993); Celeux and Govaert (1995); Day (1969); Fraley and Raftery (1998a); Marriott (1975)). This generative process is summarized by the probabilistic graphical model shown in Figure 2.1.

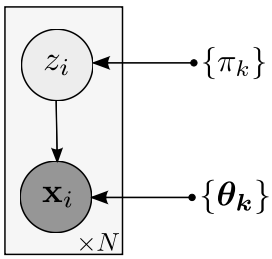


Figure 2.1: *Probabilistic graphical model for the finite mixture model.*

This thesis will focus on mixtures for multivariate real data and the Gaussian mixture which is one of the suited models to multivariate data. The Gaussian Mixture Model (GMM) has also shown a great performance in clustering applications. It is discussed in the next subsection. Several extensions, namely parsimonious ones, have been derived from the standard Gaussian mixture to accommodate more complex data, which are also considered in this thesis.

2.3 The finite Gaussian mixture model (GMM)

One of the used distributions to generate the observed data, that showed great performance in cluster analysis (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Day, 1969; Fraley and Raftery, 1998a; Ghahramani and Hinton, 1997; Marriott, 1975; McLachlan et al., 2003; McNicholas and Murphy, 2008; Scott and Symons, 1981) are the normal distributions.

Each component of this mixture model has a Gaussian density. It is parametrized by the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$ and is defined by:

$$p_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right\} \quad (2.3)$$

The Gaussian density $p_k(\mathbf{x}_i|\boldsymbol{\theta}_k)$ can be denoted as $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ or $\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Thus, the multivariate Gaussian mixture model given

as

$$p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.4}$$

is parametrized by the parameter vector $\boldsymbol{\theta} = (\pi_1, \ldots, \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$.

The generative process for the Gaussian mixture model can be similarly stated, by the two steps, as in the generative process for the general finite mixture model (Equation (2.2)). However, for the GMM case, for each component k, the observation $\mathbf{x}_i$ is generated independently from a multivariate Gaussian with the corresponding parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. This is summarized as:

$$\begin{aligned} z_i &\sim \text{Mult}(\pi_1, \ldots, \pi_k), \\ \mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} &\sim \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}). \end{aligned} \tag{2.5}$$

In the same way as for the mixture model, Figure 2.2, shows the probabilistic graphical model for the finite multivariate GMM.

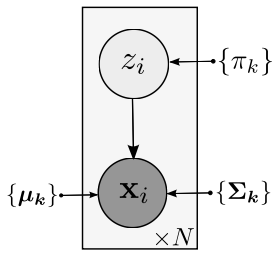


Figure 2.2: *Probabilistic graphical model for the finite GMM.*

An example of three component multivariate GMM in $\mathbb{R}^2$ with the following model parameters: $\boldsymbol{\pi} = (0.5\ 0.3\ 0.2)$, $\boldsymbol{\mu}_1 = (0.22\ 0.45)$, $\boldsymbol{\mu}_2 = (0.5\ 0.5)$, $\boldsymbol{\mu}_3 = (0.77\ 0.55)$ and $\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.018 & 0.01 \\ 0.01 & 0.011 \end{pmatrix}$, $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.011 & -0.01 \\ -0.01 & 0.018 \end{pmatrix}$, and $\boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_1$, is shown in Figure 2.3.

In modeling multivariate data, the models may suffer from the curse of dimensionality problem, causing difficulties in high-dimensional data. We refer the reader, for example, to a discussion on the curse of dimensionality problem in mixture modeling and model-based clustering in Bouveyron (2006); Bouveyron and Brunet-Saumard (2014), for further we also discuss it in the following subsection.

2.4 Dimensionality reduction and Parsimonious mixture models

One of the most important issues in modeling and clustering high-dimensional data is the curse of dimensionality. This is due to the fact that in model-

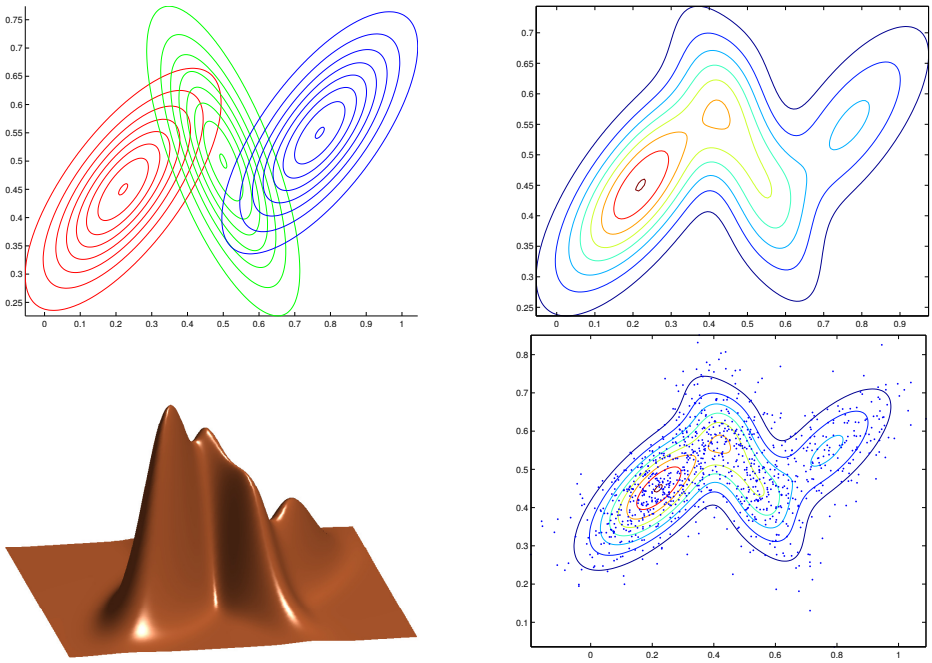


Figure 2.3: *Example of the three components multivariate GMM in* $\mathbb{R}^2$.

based clustering, an increase in the dimension, in general results an increase in the parameter space dimension. For example, for a multivariate Gaussian mixture model, with K components, the number of free parameters to estimate, for a d dimensional data is given by the following:

$$\nu(\boldsymbol{\theta}) = \nu(\boldsymbol{\pi}) + \nu(\boldsymbol{\mu}) + \nu(\boldsymbol{\Sigma}), \tag{2.6}$$

where $\nu(\boldsymbol{\pi}) = (K-1)$, $\nu(\boldsymbol{\mu}) = Kd$ and $\nu(\boldsymbol{\Sigma}) = Kd(d+1)/2$ which represent, respectively the number of mixing proportions, the mean vectors and the different values of symmetric covariance matrices. One can see in Equation (2.6), that the number of parameters to estimate for the GMM is quadratic in d, meaning that a higher-dimensional data generates a larger number of model parameters to estimate. Another issue for the Gaussian mixture model estimation arises when the number of observations n is smaller then the dimension d, this producing a singular covariance matrices, thus the model-based clustering being useless. Hopefully, the model-based clustering approaches can deal with this problem of curse of dimensionality by some approaches known in literature as: dimensionality reduction, regularization methods and parsimonious mixture models. We discuss them in the next subsections.

2.4.1 Dimensionality reduction

A first solution is to select useful characteristics from the original data, that are sufficient to represent at best the original data, that is, without no significant loss of information. For example in clustering on can cite Hall et al. (2005); Murtagh (2009).

In this formulation of dimensionality reduction different linear and non-linear data dimensionality reduction techniques are proposed for optimization of the representation space. One of the most popular approaches for dimensional reduction, the Principal Component Analysis (PCA) is a linear method firstly introduced by Hotelling (1933); Pearson (1901), or it's probabilistic version, that is Probabilistic PCA (PPCA) introduced by Tipping and Bishop (1999). We can cite also other linear dimensional reduction like Independent component analysis (ICA) (Hérault et al., 1985), Factor Analysis (FA)(Spearman, 1904), or nonlinear dimensionality reduction methods such as, Kernel Principal Component Analysis (Schölkopf et al., 1999), Relevance Feature Vector Machines (Tipping, 2001), etc.

2.4.2 Regularization methods

Another way to deal with the problem of high-dimensionality is regularization. For example, for the GMM, the issue of the curse of dimensionality is mainly related to the covariance matrix $\mathbf{\Sigma}_k$ needs to be inverted. This can be tackled with some numerical treatment namely the regularization methods, that consist in adding a numerical term to the covariance matrix before it is inversed. For example, one simple way is to add a positive term to the diagonal of the covariance matrix is given as follows:

$$\hat{\mathbf{\Sigma}}_k = \hat{\mathbf{\Sigma}}_k + \sigma_k \mathbf{I}$$

This is ridge regularization, often used in Linear Discriminant Analysis (LDA). To generalize the ridge regularization, the identity matrix can be replaced by some regularization matrix (Hastie et al., 1995). We do not focus on the regularization methods, however the reader can consider Mkhadri et al. (1997) paper for more details over the different regularization methods.

2.4.3 Parsimonious mixture models

Another way to tackle the curse of dimensionality issue are the parsimonious mixture models (Banfield and Raftery, 1993; Bensmail, 1995; Bensmail and Celeux, 1996; Celeux and Govaert, 1995; Fraley and Raftery, 1998b, 2002, 2007a,b, 2005), where the main idea is reducing the number of parameters to estimate in the mixture, by parameterising the component covariance matrices. In this work we focus on these multivariate parsimonious Gaussian mixture models for modeling and clustering high-dimensional data.

Constrained Gaussian Mixture Models One of traditional way that introduces the parsimonious Gaussian models reducing the number of parameters to estimate is to consider constraints for the covariance matrix. The most frequent used constraints for Gaussian mixture models are listed as follows:

1. the GMM itself consisting of the full covariance matrices $\mathbf{\Sigma}_k$, for all the components $\forall k = 1 \dots K$, which is abbreviated of Full-GMM,
2. the Com-GMM assume that the Gaussian mixture model consists of components with equal covariance matrices $\mathbf{\Sigma}_k = \mathbf{\Sigma}$, $\forall k = 1 \dots K$.
3. the Diag-GMM in which all the components have diagonal covariance matrices: $\mathbf{\Sigma}_k = \text{diag}(\sigma_{k1}^2, \dots, \sigma_{kd}^2)$,
4. the Com-Diag-GMM model, have a common diagonal covariance for all components $\forall k = 1 \dots K$ of the model: $\mathbf{\Sigma}_k = \mathbf{\Sigma} = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$,
5. the Sphe-GMM suppose spherical covariances for all the components $\forall k = 1 \dots K$ of the model: $\mathbf{\Sigma}_k = \sigma_k^2 \mathbf{I}$,
6. the Com-Sphe-GMM model is a spherical model with equal covariances, for all the components $\forall k = 1 \dots K$, that is: $\mathbf{\Sigma}_k = \mathbf{\Sigma} = \sigma^2 \mathbf{I}$.

The number of mixture parameters related to the covariance matrices, for these six constrained GMMs, is summarized in Table 2.1.

Constrained GMM	$\nu(\mathbf{\Sigma})$
Full-GMM	$Kd(d+1)/2$
Com-GMM	$d(d+1)/2$
Diag-GMM	Kd
Com-Diag-GMM	d
Sphe-GMM	K
Com-Sphe-GMM	1

Table 2.1: *The constrained Gaussian Mixture Models and the corresponding number of free parameters related to the covariance matrix.*

To illustrate the effect of the constraints on the model dimension, consider the Full-GMM and the Com-GMM with equal number of components $K = 3$. Figure 2.4 shows the number of free parameters $\nu(\boldsymbol{\theta})$ as function of the data dimension. One can see that, the number of free parameters to estimate for the general Full-GMM gets significant larger, than the constraint Com-GMM, as the data dimension grows. We refer the reader on the paper of Bouveyron and Brunet-Saumard (2014); McNicholas and Murphy (2008) for more detailed description of these constrained models.

Parsimonious mixture models via eigenvalue decomposition of the covariance matrix A similar way of extending the finite GMM to parsimonious GMM (PGMM), (Banfield and Raftery, 1993; Celeux and Govaert,

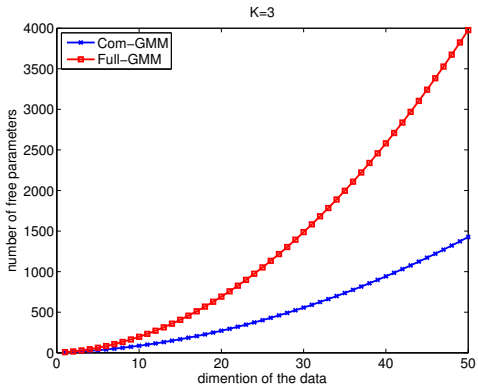


Figure 2.4: *The number of parameters to estimate for the Full-GMM and the Com-GMM in respect of the dimension of the data and the number of components $K = 3$.*

1995) consists in exploiting an eigenvalue decomposition of the group covariance matrices, which provides a wide range of very flexible models with different clustering criteria. The group covariance matrix $\mathbf{\Sigma}_k$ for each cluster k, in these parsimonious models, is decomposed as

$$\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{2.7}$$

where the scalar $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$ determines the volume of cluster k, $\mathbf{D}_k$ is an orthogonal matrix of eigenvectors of $\mathbf{\Sigma}_k$ determines the orientation and $\mathbf{A}_k$ that is the shape of cluster k, is a diagonal matrix with determinant 1 whose diagonal elements are the normalized eigenvalues of $\mathbf{\Sigma}_k$ in a decreasing order (Celeux and Govaert, 1995). This decomposition leads to several flexible models, going from the simplest spherical models, to the complex general one, and hence is adapted to various clustering situations. Table 2.2 enumerates the 14 parsimonious GMMs that can be obtained by the decomposition (2.7). They are implemented in the MCLUST software Fraley and Raftery (1998b, 2007b). Notice that their names consists of three different letters E, V and I that encodes the geometric characteristics: volume, orientation and shape. The letter E means equal, V means varying across components and clusters, and I refers to the identity matrix specifying the shape or orientation. Giving an example we may refer to a VEI model where the volume clusters may vary (V), the shape of the clusters are equal (E), and the orientation is the identity (I). Indeed this model refers to the diagonal model $\lambda_k \mathbf{A}$. For example, the Full-GMM model corresponding to the $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ decomposition is named VVV since it has varying volume, shape and orientation. Note that the models flagged with the star in Table 2.2 are not available in the MCLUST application.

Also one can see that Table 2.2 distinguishes between three different

Model	Name	Number of free parameters
$\lambda\mathbf{I}$	EII	$\upsilon + 1$
$\lambda_k\mathbf{I}$	VII	$\upsilon + d$
$\lambda\mathbf{A}$	EEI	$\upsilon + d$
$\lambda_k\mathbf{A}$	VEI	$\upsilon + d + K - 1$
$\lambda\mathbf{A}_k$	EVI	$\upsilon + Kd - K + 1$
$\lambda_k\mathbf{A}_k$	VVI	$\upsilon + Kd$
$\lambda\mathbf{DAD}^T$	EEE	$\upsilon + \omega$
$\lambda_k\mathbf{DAD}^T$	VEE*	$\upsilon + \omega + K - 1$
$\lambda\mathbf{DA}_k\mathbf{D}^T$	EVE*	$\upsilon + \omega + (K-1)(d-1)$
$\lambda_k\mathbf{DA}_k\mathbf{D}^T$	VEE*	$\upsilon + \omega + (K-1)d$
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	EEV	$\upsilon + K\omega - (K-1)d$
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	VEV	$\upsilon + K\omega - (K-1)(d-1)$
$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	EVV*	$\upsilon + K\omega - (K-1)$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	VVV	$\upsilon + K\omega$

Table 2.2: *The Parsimonious Gaussian Mixture Models via eigenvalue decomposition, the model names as in the MCLUST software, and the corresponding number of free parameters* $\upsilon = \nu(\boldsymbol{\pi}) + \nu(\boldsymbol{\mu}) = (K-1) + Kd$ *and* $\omega = d(d+1)/2$, K *being the number of mixture components and* d *the number of variables for each individual.*

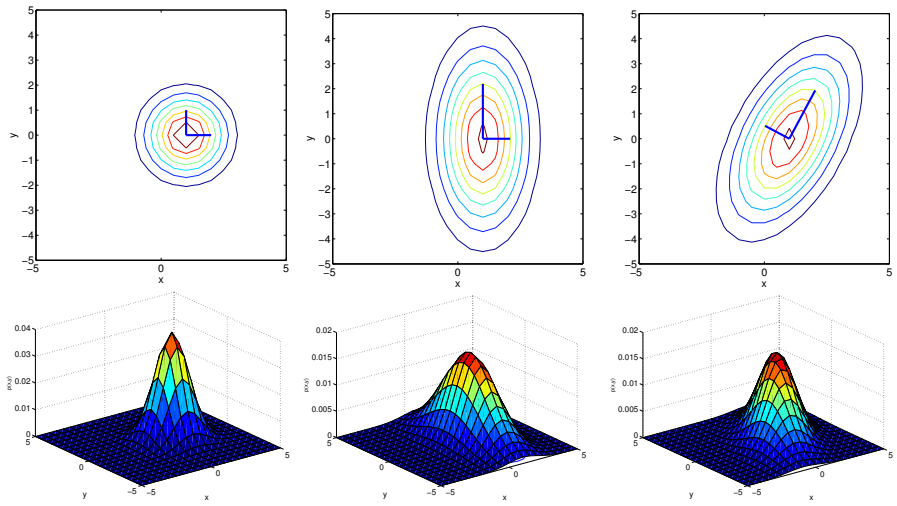
families, that are the spherical family, the diagonal family, and the general family.

Figure 2.6 illustrates the geometrical representation of all the fourteen possible parsimonious models, issued from the decomposition (2.7) of the covariance matrix. One can see how the volume, orientation and the shape can vary between all 14 models.

These models will consist the bases of our contributions. Later, we will provide both the Bayesian parametric formulation, as well as the full Bayesian non-parametric derivations.

In model-based clustering using GMMs, the model parameters are usually estimated into a maximum likelihood estimation (MLE) framework by maximizing the observed data likelihood. This is usually performed by the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008) or EM extensions (McLachlan and Krishnan, 2008), such as the CEM algorithm (Celeux and Govaert, 1992, 1995; Samé et al., 2007), or stochastic EM version as in Celeux and Diebolt (1985); Celeux et al. (1995, 1996).

In the next section, we describe the maximum likelihood (ML) fitting of the finite mixture, using the EM algorithm, and focusing on the GMM and parsimonious GMMs.



(a) *Spherical* (b) *Diagonal* (c) *General*

Figure 2.5: *2D Gaussian plots of a spherical, diagonal and full covariance matrix, representing all three families of the parsimonious GMM.*

2.5 Maximum likelihood (ML) fitting of finite mixture models

The model parameters $\boldsymbol{\theta}$ are estimated from an i.i.d dataset $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. For example, for the multivariate GMM, the parameter vector to be estimated is $\boldsymbol{\theta} = (\pi_1, \ldots \pi_K, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$. One of the main framework that is used for estimation of these model parameters are the Maximum Likelihood (MLE) framework (Banfield and Raftery, 1993; McLachlan and Basford, 1988; McLachlan and Krishnan, 2008; Samé et al., 2007). In this framework, the model parameters $\boldsymbol{\theta}$ are estimated by maximizing the following observed data log-likelihood.

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{2.8}$$

This log-likelihood can not be maximized in analytic way. The standard way, to do this, is to do it iteratively, via the EM algorithm. The complete data log-likelihood, needed to derive the EM where the complete data $(\mathbf{X}, \mathbf{z})$, $\mathbf{z}$ being the allocation variables, with z_i the label of the component generating the observation $\mathbf{x}_i$, is given by:

$$\log L_c(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.9}$$

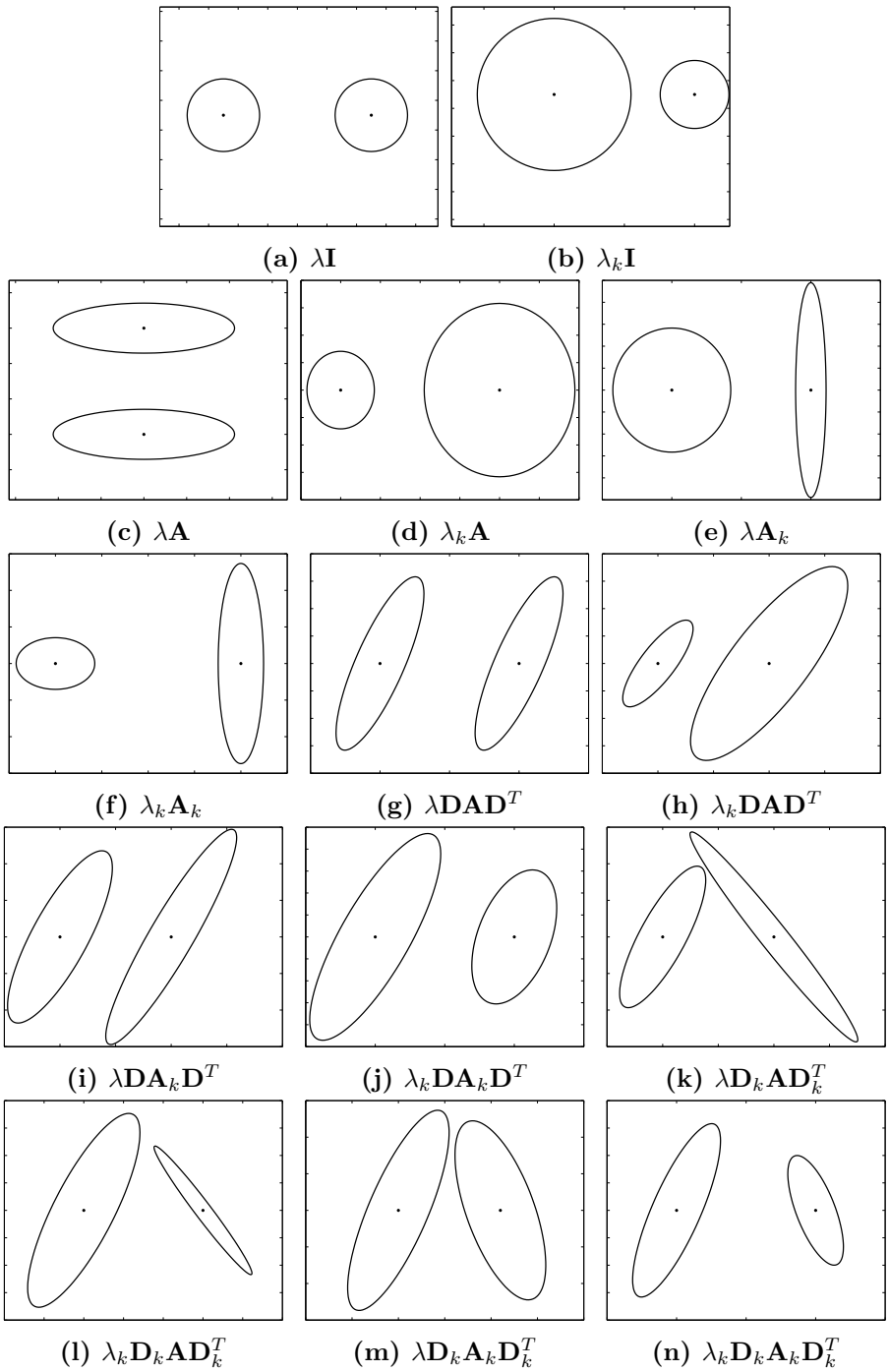


Figure 2.6: *The geometrical representation of the* 14 *parsimonious Gaussian mixture models with the eigenvalue decomposition (2.7).*

where z_{ik} are indicator variables such that $z_{ik} = 1$ if $z_i = k$ and $z_{ik} = 0$ otherwise.

2.5.1 ML fitting via the EM algorithm

The maximum likelihood estimation framework is usually performed by the Expectation-Maximization (EM) algorithm (Dempster et al., 1977; McLachlan and Krishnan, 2008). The EM for the finite GMM is recalled in the following.

Suppose, the initial vector parameters values for the GMM are given by $\boldsymbol{\theta}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_K^{(0)})$. The Expectation-Maximization (EM) clustering algorithm is an iterative algorithm, that consists of two main steps: the Expectation E-step and the Maximization M-step.

E-Step First, the E-step, computes the expectation of the complete data log-likelihood (2.9) given the observations $\mathbf{X}$ and the current value of the model parameters vector $(\boldsymbol{\theta}^{(t)})$, (t) being the current iteration number. This conditional expectation is known as the $\mathcal{Q}$-function:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) &= E[\log L_c(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})|\mathbf{X}; \boldsymbol{\theta}^{(t)}] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} E[z_{ik}|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}] \log \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} p(z_{ik} = 1|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}) \log \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \log \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad (2.10)
\end{aligned}$$

where

$$\tau_{ik}^{(t)} = p(z_{ik} = 1|\mathbf{X}; \boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}, \qquad (2.11)$$

is the posterior probability that $\mathbf{x}_i$ is generated from the kth component density.

M-Step The M-step consists in updating the parameter vector $\boldsymbol{\theta}$ by maximizing the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$, that is

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}). \qquad (2.12)$$

The parameter vector update in the GMM (see for example McLachlan and Krishnan (2008); Redner and Walker (1984)) are given by:

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ik}^{(t)}, \tag{2.13}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{n_k^{(t)}}\sum_{i=1}^{n}\tau_{ik}^{(t)}\mathbf{x}_i, \tag{2.14}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\mathbf{W}_k^{(t)}}{n_k^{(t)}}, \tag{2.15}$$

where

$$n_k^{(t)} = \sum_{i=1}^{n}\tau_{ik}^{(t)}, \tag{2.16}$$

is the expected number of observations that belong to the kth component. and $\mathbf{W}_k^{(t+1)}$ is the expected scattering matrix of kth component given by:

$$\mathbf{W}_k^{(t+1)} = \sum_{i=1}^{n}\tau_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T \tag{2.17}$$

EM initialization One of the crucial steps in EM algorithm is the initialization step, because that EM maximizes locally the log-likelihood. Therefore the quality of the estimation and the speed of the convergence depends directly on the initialization step. To solve this issue some methods where discussed in the literature, in particular Biernacki (2004). One of the most used method, is running the EM algorithm many times with different initializations, and then the maximum log-likelihood solution of those runs to be selected. The EM algorithm initializations can be done with:

- random initialization,
- by computing the initial parameter vector by other clustering algorithms like K-means (MacQueen, 1967) , one of the EM extensions (McLachlan and Krishnan, 2008) like the Classification EM (Celeux and Diebolt, 1985), Stochastic EM (Celeux and Govaert, 1992), etc,
- initialization by some EM steps itself.

For future discussion on the subject, the reader is referred to Biernacki et al. (2003); Biernacki (2004).

EM stopping rule One of the main properties of the EM algorithm is that the likelihood must increment in each step (McLachlan and Krishnan, 2008; Neal and Hinton, 1998; Wu, 1983). So the convergence, can be supposed

to be reached when the log-likelihood improvement from one iteration to another is less then a prefixed threshold, that is:

$$\left| \frac{\log L(\boldsymbol{\theta})^{(t+1)} - \log L(\boldsymbol{\theta})^{(t)}}{\log L(\boldsymbol{\theta})^{(t)}} \right| \leq \epsilon.$$

The Pseudo-code 1 summarizes the Expectation-Maximization algorithm for ML fitting of the GMM.

Algorithm 1 Expectation-Maximization via ML estimation for Gaussian Mixture Models

Inputs: Data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, # of mixture components K

Fix threshold $\epsilon > 0$ $t \leftarrow 0$
Initialize $\boldsymbol{\theta}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_K^{(0)})$
while increment in log-likelihood $> \epsilon$ **do**
  E-Step
  for $k \leftarrow 1$ **to** K and $i \leftarrow 1$ **to** n **do**
    Compute $\tau_{ik}^{(t)}$ using Equation (2.11).
  end for
  M-Step
  for $k \leftarrow 1$ **to** K **do**
    Compute $\pi_k^{(t+1)}$ using Equation (2.13)
    Compute $\boldsymbol{\mu}_k^{(t+1)}$ using Equation (2.14)
    Compute $\boldsymbol{\Sigma}_k^{(t+1)}$ using Equation (2.15)
  end for
  $t \leftarrow t + 1$
end while

Outputs: The Gaussian parameter vector $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t)}$ and the fuzzy partition of the data $\hat{\tau}_{ik} = \tau_{ik}^{(t)}$

Once the GMM model parameters $\hat{\boldsymbol{\theta}}_{ML}$ are estimated, a partition of the data into K clusters can then be obtained by maximizing the posterior component probabilities $\hat{\tau}_{ik}$, that is, by computing the cluster labels:

$$\hat{z}_i = \arg \max_{1<k<K} \hat{\tau}_{ik}. \tag{2.18}$$

2.5.2 Illustration of ML fitting of a GMM

To illustrate the EM, we consider the well-known bivariate Old Faithful Geyser dataset (Azzalini and Bowman, 1990) composed of $n = 252$ observations in $\mathbb{R}^2$ shown in Figure 2.7. Note that a normalization pre-processing step was performed. The GMM partition, as well as the mixture component ellipse densities, obtained by the EM algorithm, and the stored log-likelihood

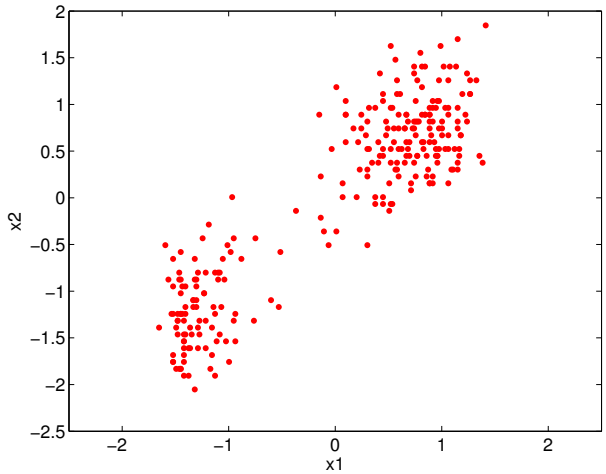


Figure 2.7: *Old Faithful Geyser data set.*

values for each EM step are shown in Figure 2.8. The mixture model, with two Gaussian components is learned with the EM algorithm. The initialization of the model parameters was made by K-means algorithm (MacQueen, 1967). We used two components, as several model-based clustering methods, in the literature, that infer two components for this dataset.

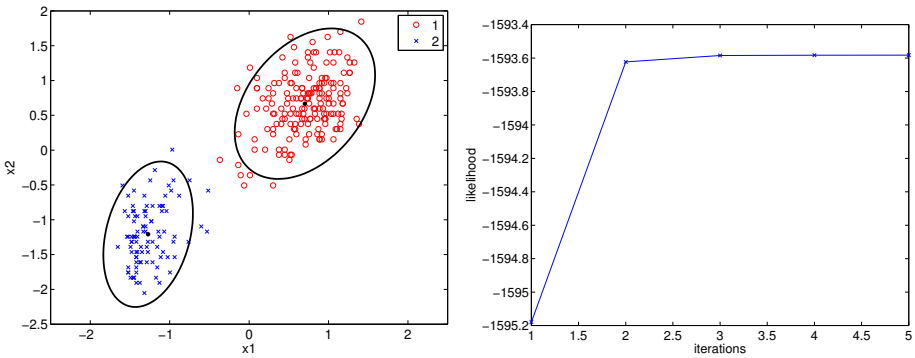


Figure 2.8: *GMM clustering with the EM algorithm for the Old Faithful Geyser. The obtained partition (left) and the log-likelihood values at each EM iteration (right).*

We also give an illustrative example for clustering the Iris data set studied by Fisher (1936). The Iris dataset contains $n = 150$ samples of Iris flowers covering three Iris species: setosa, virginica and versicolor, that is $K = 3$, with 50 samples for each specie. Four features were measured for each sample ($d = 4$): the length and the width of the sepals and petals, in centimetres. Figure 2.9 shows the true partition of the Iris data set in the space of the components 3 (petal length) and 4 (petal width).

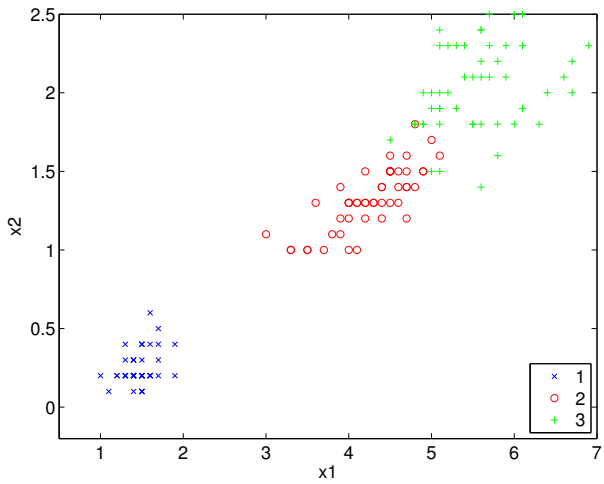


Figure 2.9: *Iris data set in the space of the components 3 (x1: petal length) and 4 (x2: petal width)*

We cluster the data set by learning a three components GMM with the EM algorithm. The obtained partition as well as the density ellipses and the log-likelihood for each of the EM step are given in Figure 2.10.

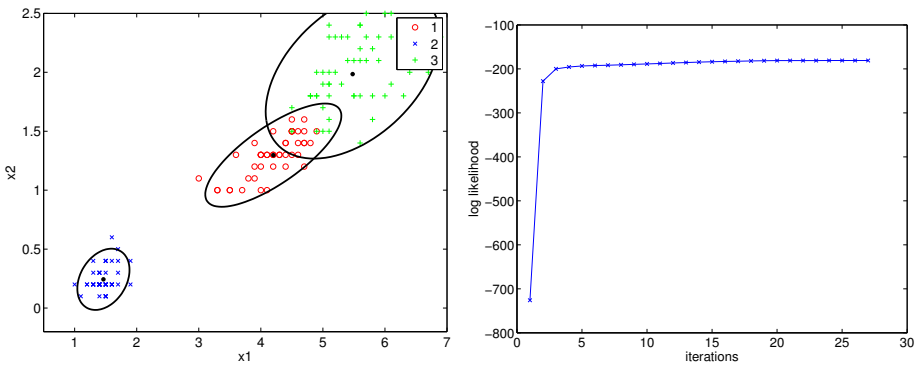


Figure 2.10: *Iris data set clustering by applying the EM algorithm for the GMM, with the obtained partition and the ellipse densities (left) and the log-likelihood values at each iteration (right).*

2.5.3 ML fitting of the parsimonious GMMs

Celeux and Govaert (1995) introduces the parsimonious Gaussian mixture by the eigenvalue decomposition of the covariance matrices, which provides 14 different models given in Table 2.2. These 14 models can be estimated by the EM clustering algorithm.

The EM scheme for the parsimonious models is as follows. The eigen-

value decomposition of the covariance model can be choosen a priori and is given as an input by the user. The E-step of the EM algorithm outlined in Pseudo-code 1 does not change. However, because parsimonious Gaussian mixture models vary by the eigenvalue decomposition of the covariance matrix for each cluster, the derivation of the M-step is computed according to it. As a result we have the same estimation of the mixture proportions (Equation (2.13)) and the mean vectors (Equation (2.14)). However, the covariance matrix is estimated according to it's chosen decomposition. More details on the M-step for the ML fitting of the parsimonious GMMs can be found in Bensmail and Celeux (1996); Celeux and Govaert (1995).

As EM maximizes locally the likelihood, the initialization step of the EM remains always one of the crucial steps that can produce not a satisfactory output. Therefore it is consigned to make the initialization as possible near to the expected parameter values. A restriction of each of the eigenvalue decomposition models, given in Table 2.2, is considered for the initialization step. For instance, the spherical model $\lambda_k\mathbf{I}$ have the spherical initialization where the volume of the cluster varies between clusters.

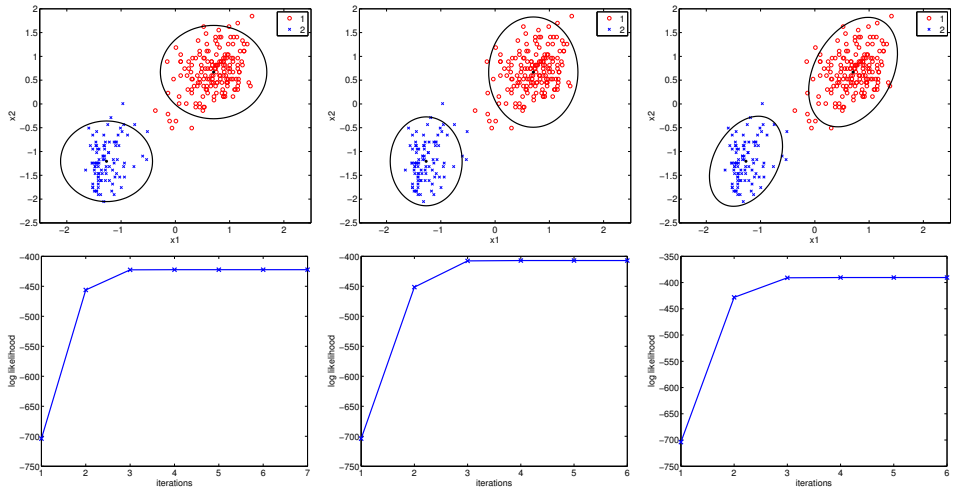
2.5.4 Illustration: ML fitting of parsimonious GMMs

To illustrate the EM algorithm for the parsimonious Gaussian mixture models we first investigate three different family of models (spherical, diagonal and general) by varying the cluster volume while the orientation and the shape remain unchanged for all clusters.

First, we apply the parsimonious GMM with the EM algorithm on the Old Faithful Geyser data set for illustration. We used two Gaussian components ($K = 2$) for this dataset. We considered three parsimonious GMM models, which are the spherical model $\lambda_k\mathbf{I}$, the diagonal model $\lambda_k\mathbf{A}$ and the general model $\lambda_k\mathbf{DAD}^T$. These models are considered so that the clusters have different volume, but equal orientation and shape. Figure 2.11 shows the obtained partitions, the component ellipse densities, as well as the log-likelihood values for the EM iterations.

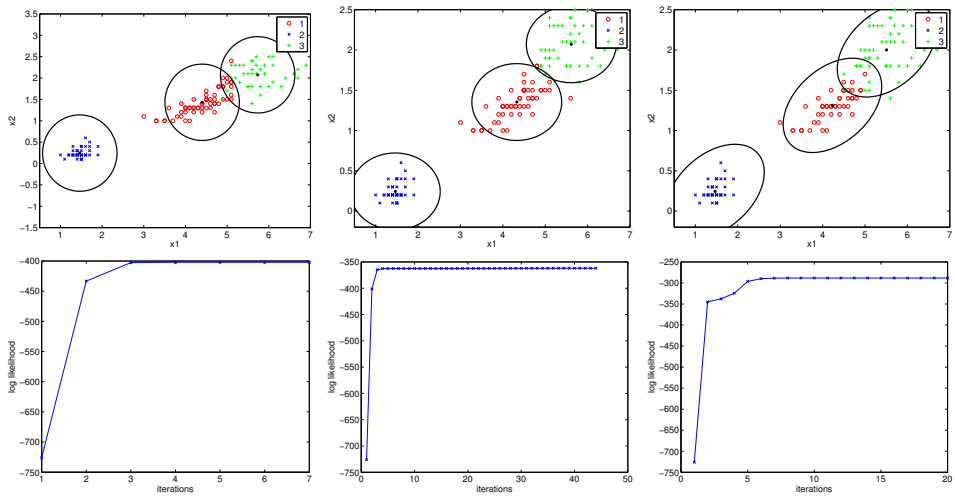
Now we apply the parsimonious GMM with the EM algorithm on the Iris data. We consider three other models, which are the spherical model $\lambda\mathbf{I}$, the diagonal model $\lambda\mathbf{A}$ and the general model $\lambda\mathbf{DAD}^T$. These models are constrained so that the clusters have the same volume, orientation and shape. Figure 2.12 shows the obtained partitions, the component ellipse densities, as well as the log-likelihood values during the EM iterations.

In the next section, we discuss the model selection and comparison in the parametric mixture models. This answers the problem of selecting the number of mixture components. For the parsimonious models, the additional feature, that is the choosing of the models structure is also investigated.



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

Figure 2.11: *Clustering the Old Faithful Geyser data set with the EM algorithm for the Parsimonious GMM. The obtained partition and the ellipse densities (top) and the log-likelihood values for each EM step (bottom). The spherical model* $\lambda_k\mathbf{I}$ *(left), the diagonal family model* $\lambda_k\mathbf{A}$ *(middle) and the general model* $\lambda_k\mathbf{DAD}^T$ *(right).*



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

Figure 2.12: *Clustering the Iris data set with the EM algorithm for the Parsimonious GMM. The obtained partition and the ellipse densities (top) and the log-likelihood values for each EM step (bottom). The spherical model* $\lambda\mathbf{I}$ *(left), the diagonal family model* $\lambda\mathbf{A}$ *(middle) and the general model* $\lambda\mathbf{DAD}^T$ *(right).*

2.6 Model selection and comparison in finite mixture models

The number of mixture components is usually assumed to be known for the parametric model-based clustering approaches. Another issue in the finite mixture model-based clustering approach is therefore the one of selecting the optimal number of mixture components. This problem, generally called model selection, is in general performed through a two-fold strategy by selecting the best model from pre-established inferred model candidates. The selection task is made by choosing a model from a set of possible models, that fits at best the data, and thus in the sense of a model selection criterion. Notice that, for the parsimonious models, which have different structures, the model selection contains an additional feature, that is the one of choosing the best model structure (i.e., the decomposition of the covariance matrix $\boldsymbol{\Sigma}_k$).

A common way for model selection, is to use an overall score function that is represented by two terms. The first one represents the goodness of the specified model (how well the selected model fits the data), and the second one, is a penalty term that governs the model complexity. In consequence, the model selection procedure in general aims at minimizing the following score function:

$$\text{score(model)} = \text{error(model)} + \text{penalty(model)}. \tag{2.19}$$

The complexity of some model $\mathcal{M}$ being directly related to the number of it's free parameters $\nu(\boldsymbol{\theta})$

Letting $\{\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_M\}$ be a set of considered models from which we wish to choose the best one. The choice of the optimal model can be performed via penalized log-likelihood criteria such as the Bayesian Information Criterion (BIC) (Schwarz, 1978), the Akaike Information Criterion (AIC) (Akaike, 1974), AIC3 (Bozdogan, 1983), the Approximate Weight of Evidence (AWE) criterion (Banfield and Raftery, 1993), or the Integrated Classification Likelihood criterion (ICL) (Biernacki et al., 2000), etc. More information on the model selection with information criteria, see for example Biernacki (1997); Biernacki and Govaert (1998); Claeskens and Hjort (2008); Konishi and Kitagawa (2008). In this work, we consider some of them, which are widely used in the literature.

2.6.1 Model selection via information criteria

Assume that the model $\mathcal{M}_1$ is parametrized by the parameter vector $\boldsymbol{\theta}_m$. $\hat{\boldsymbol{\theta}}_m$ is the maximum likelihood estimator (respectively the maximum complete likelihood estimator of $\boldsymbol{\theta}_m$). The most used information criteria for model selection are the Akaike Information Criteria (AIC) (Akaike, 1974), the AIC3

(Bozdogan, 1983), the Bayesian Information Criteria (BIC) (Schwarz, 1978), the Integrated Classification Likelihood (ICL) (Biernacki et al., 2000), and the Approximate Weight of Evidence (AWE) (Banfield and Raftery, 1993). They are respectively defined as:

$$
\begin{aligned}
AIC(\mathcal{M}_m) &= \log L(\mathbf{X}|\hat{\boldsymbol{\theta}}_m) - \nu_m, && (2.20)\\
AIC3(\mathcal{M}_m) &= \log L(\mathbf{X}|\hat{\boldsymbol{\theta}}_m) - \frac{3\nu_m}{2}, && (2.21)\\
BIC(\mathcal{M}_m) &= \log L(\mathbf{X}|\hat{\boldsymbol{\theta}}_m) - \frac{\nu_m \log(n)}{2}, && (2.22)\\
ICL(\mathcal{M}_m) &= \log L_c(\mathbf{X}, \mathbf{z}|\hat{\boldsymbol{\theta}}_m) - \frac{\nu_m \log(n)}{2}, && (2.23)\\
AWE(\mathcal{M}_m) &= \log L_c(\mathbf{X}, \mathbf{z}|\hat{\boldsymbol{\theta}}_m) - (\nu_m(\frac{3}{2} + log(n))). && (2.24)
\end{aligned}
$$

where $\log L(\mathbf{X}|\hat{\boldsymbol{\theta}}_m)$ is the maximum value of the observed data log-likelihood and $\log L_c(\mathbf{X}, \mathbf{z}|\hat{\boldsymbol{\theta}}_m)$ is the maximum value of the complete data log-likelihood.

These information criteria, can also be seen as approximations of the Bayes Factor (Fraley and Raftery, 1998a; Kass and Raftery, 1995). Because Bayes Factor is considered a fully Bayesian method form model selection and comparison between models, we will be discussed it in Chapter 3 and Chapter 4.

For the parsimonious models, the model selection answers not just to the question: "how much clusters (components) are in the data?", but also allows to provide the best model structure (Fraley and Raftery, 1998a). The strategy for the parsimonious finite mixture models regarding the estimation of the number of clusters and the best model structure is investigated in this work.

2.6.2 Model selection for parsimonious GMMs

For the parsimonious finite Gaussian mixture models, the model selection task can be separated into two issues to investigate. First, the selection of components number (i.e. clusters K) in the mixture, and second, what parsimonious model fits at best the data. Let $K_{\max}$ be the maximum number of components in the mixture and $(\mathcal{M}_1, \ldots, \mathcal{M}_M)$ a set of parsimonious Gaussian mixture models with different eigenvalue decomposition of the covariance matrix. We derived the Pseudo-code 2 for the model selection strategy of the parsimonious GMMs that was found to be effective in the literature (Dasgupta and Raftery, 1998; Fraley and Raftery, 1998a, 2007a, 2005).

Thus the number of mixture components (classes) and the the eigenvalue decomposition of the covariance matrix that fit at best the data are determined in one run.

Algorithm 2 Model selection for parsimonious Gaussian mixture models

Inputs: $K_{\max}$, specified model structure $(\mathcal{M}_1, \ldots, \mathcal{M}_M)$.

```
for k ← 1 to K_max do
    for m ← 1 to M do
        Compute the MLE θ̂_km (e.g. via EM);
        Compute IC(θ̂_km) where IC(θ̂_km) is the Information Criterion value
        given the estimated model parameters θ̂_km for model structure m
        and k components (e.g. for BIC (2.22)).
    end for
end for
Choose the model having the highest information criterion value M̂
```

Outputs: The selected model $\hat{\mathcal{M}}$

2.6.3 Illustration: Model selection and comparison via information criteria

We consider the Old Faithful Geyser and Iris datasets to investigate the model selection for six parsimonious Gaussian mixture models, that are, two models from each family: $\lambda\mathbf{I}$ and $\lambda_k\mathbf{I}$ for the spherical case, $\lambda\mathbf{A}$ and $\lambda_k\mathbf{A}$ for the diagonal case, and $\lambda_k\mathbf{DAD}^T$ and $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ for the general case. The EM algorithm is used and initialized by K-means. The BIC (2.22), ICL (2.23) and AWE (2.24) criteria are computed for this model selection experiment.

The top plot of Figure 2.13 illustrates the model selection for the Old Faithful Geyser dataset.

The BIC criterion selects: 5 clusters for the spherical models and therefore overestimates the number of clusters, 4 clusters for the diagonal model, which has different cluster volume, that is, $\lambda_k\mathbf{A}$, 3 clusters for the diagonal model, which has equal cluster volume $\lambda\mathbf{A}$, and for the general model, which has different cluster volume $\lambda_k\mathbf{DAD}^T$, 2 clusters for the Full-GMM model. The highest BIC criterion value, that selects the best model, was obtained by the $\lambda_k\mathbf{DAD}^T$ model.

The ICL criterion selects: 4 clusters for the spherical model, which has different cluster volume $\lambda_k\mathbf{I}$, therefore overestimating the number of clusters, 3 clusters for the spherical model, which has equal cluster volume $\lambda\mathbf{I}$, 2 clusters for the rest of the model candidates. The highest ICL criterion value, that selects the best model, was obtained by the Full-GMM, that is $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ model.

Finally, the AWE criterion is investigated. One can see that, for this dataset, the AWE criteria does not overestimates the number of components for the model candidates. AWE criterion selects 3 clusters for the diagonal model $\lambda\mathbf{A}$, while for the rest of the models 2 clusters are selected. The highest AWE criterion value, that selects the best model, was obtained by the

$\lambda_k\mathbf{DAD}^T$ model. Highlight, that in Figure 2.13, the descending values for the studied information criterion, the AWE criteria descends more sharply then the BIC and ICL criteria meaning a more decisive model selection.

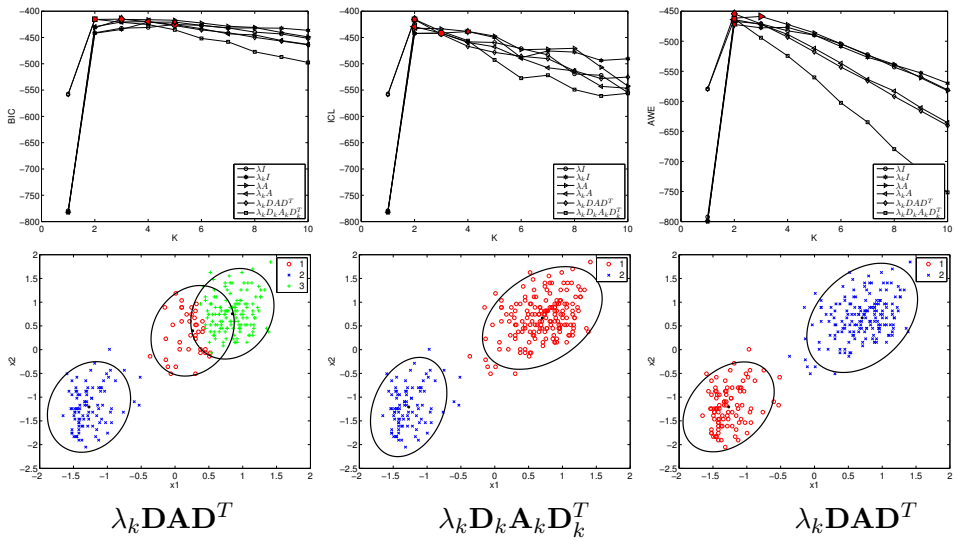


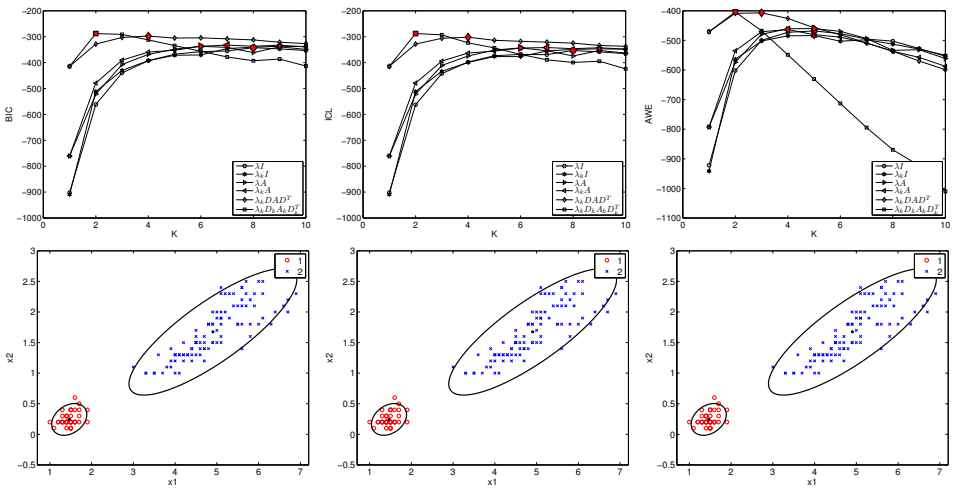
Figure 2.13: *Model selection for Old Faithful Geyser dataset with BIC (left), ICL (middle) and AWE (right). The top plot shows the value of the IC for different models and different mixture components* ($k = 1, \ldots, 10$). *The bottom plot show the selected model partition and the corresponding mixture component ellipse densities.*

The top plot of Figure 2.13 illustrates the model selection for the Iris dataset. The BIC, ICL and AWE criterion are investigated. For all of these information criterion, the highest value that selects the best model was the Full-GMM model. However, we can see that the AWE criterion selects the true number of clusters equal to 3, for the general model, that is, $\lambda_k\mathbf{DAD}^T$.

2.7 Conclusion

In this chapter, we presented state of the art approach on mixture modeling for model-based clustering. We focused on the Gaussian case and the parsimonious mixture models. We discussed the use of the EM algorithm which constitutes the essential feature for model fitting. Then we showed how the model selection and comparison can be performed in this ML fitting framework.

In the next chapter, we will address the problem of model-based clustering from a Bayesian prospective and implement several alternative Bayesian parsimonious mixtures for clustering.



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

Figure 2.14: *Model selection for Iris dataset with BIC (left), ICL (middle) and AWE (right). The top plot shows the value of the IC for different models and different mixture components* ($k = 1, \ldots, 10$)*. The bottom plot show the selected model partition and the corresponding mixture component ellipse densities.*

- Chapter 3 -

Bayesian mixture models for model-based clustering

Contents

3.1 Introduction

In this chapter, we investigate the mixture models in a Bayesian framework, rather than a ML fitting, as described in Chapter 2. After an account on Bayesian mixture modeling, we focus on Bayes formulation of the previously described parsimonious Gaussian mixtures. We present the Maximum A Posteriori estimation using in particular Markov Chain Monte Carlo sampling. The model selection and comparison is addressed from a Bayesian point of view by using Bayes Factors. Gibbs sampling technique is implemented for the various parsimonious GMMs, which we apply and assess in different simulation scenarios.

3.2 The Bayesian finite mixture model

Described earlier in Chapter 2, the parametric model-based clustering have shown great performances in density estimation and model-based clustering (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Day, 1969; Fraley and Raftery, 1998a; Marriott, 1975; Scott and Symons, 1981). However, a first issue for the ML parameter estimation of the mixture models, is that it may fail due to singularities or degeneracies as highlighted in Fraley and Raftery (2007a, 2005); Ormoneit and Tresp (1998); Snoussi and Mohammad-Djafari (2000, 2005); Stephens (1997).

The Bayesian formulation of the finite mixture models allows to avoid these problems by replacing the MLE by the maximum a posterior (MAP) estimator. This is namely achieved by basically giving some penalization term, namely regularization, to the observed data likelihood function. The estimation of the Bayesian mixtures via the posterior simulations goes back to Evans et al. (1992); Gelman and King (1990); Verdinelli and Wasserman (1991). The Bayesian estimation methods for mixture models have lead to intensive research in the field for dealing with the problems encountered in MLE for mixtures. One can cite for example the following papers on the subject: Bensmail and Meulman (2003); Bensmail et al. (1997); Diebolt and Robert (1994); Escobar and West (1994); Gelman et al. (2003); Marin et al. (2005); Richardson and Green (1997); Robert (1994); Stephens (1997). Bayesian approaches allow to avoid these problems by replacing the MLE by the maximum a posterior (MAP) estimator.

Suppose the mixture model, given in Equation (2.1), with parameters $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$. The Bayesian mixture model incorporates prior distribution on these parameters. In this thesis we focus on conjugate priors, for which the posterior are easy to derive. The generative process of the Bayesian mixture models is given as follows.

The first step is to sample the model parameters from the prior, that is,

for example, to sample the mixing proportions from their conjugate Dirichlet prior distribution. The parameters $\boldsymbol{\theta}_k$ are sampled according to a prior base distribution noted G_0. This can be summarized as follows:

$$\begin{aligned} \boldsymbol{\pi}|\boldsymbol{\alpha} &\sim \text{Dir}\left(\tfrac{\alpha_1}{K}, \ldots, \tfrac{\alpha_K}{K}\right), \\ z_i|\boldsymbol{\pi} &\sim \text{Mult}(1; \pi_1, \ldots, \pi_k), \\ \boldsymbol{\theta}_{z_i}|G_0 &\sim G_0, \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim p_k(\mathbf{x}_i|\boldsymbol{\theta}_{z_i}). \end{aligned} \tag{3.1}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, the concentration hyperparameters of the Dirichlet prior distribution, $p_k(\mathbf{x}_i|\boldsymbol{\theta}_{z_i})$ is a conditional component density function with parameter $\boldsymbol{\theta}_{z_i}$. The labels z_i are sampled according to multinomial distribution with parameters being the mixing proportions $\boldsymbol{\pi}$, which are sampled according to the Dirichlet distribution. The probabilistic graphical model for the finite Bayesian mixture model is shown in Figure 3.1.

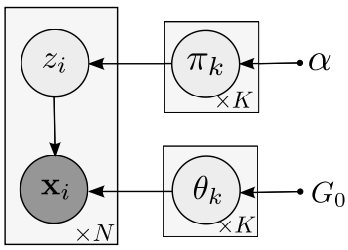


Figure 3.1: *Probabilistic graphical model for the Bayesian mixture model.*

In the next section, we discuss the Bayesian mixture model when data is considered to be Gaussian distributed.

3.3 The Bayesian Gaussian mixture model

The Bayesian GMM is also one of the most successful and popular models in the literature. It has also shown great performances in density estimation and cluster analysis. For additional to review on Bayesian GMMs, we refer the reader to the following key papers: Bensmail et al. (1997); Diebolt and Robert (1994); Fraley and Raftery (2007a, 2005); Ormoneit and Tresp (1998); Richardson and Green (1997); Robert (1994); Snoussi and Mohammad-Djafari (2000); Stephens (1997, 2000).

The generative process for the Bayesian GMM is given by Equation (2.3), where the parameters and the priors are those corresponding to the Gaussian case. Using conjugate priors[1] is commonly used in Bayesian mixture models. For the GMM case, the Gaussian parameter model priors are a

[1] In Bayesian statistics if the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ is in the same family as the prior distribution $p(\boldsymbol{\theta})$, than this prior is considered to be a conjugate distribution.

multivariate Normal distribution for the mean vector parameter $\boldsymbol{\mu}_k$ and an inverse-Wishart distribution for the covariance matrix $\boldsymbol{\Sigma}_k$. Thus, the base measure, G_0, from Equation (3.1), corresponds to the following prior:

$$\begin{aligned} \boldsymbol{\Sigma}_{z_i} &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0), \\ \boldsymbol{\mu}_{z_i}|\boldsymbol{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \tfrac{\boldsymbol{\Sigma}_k}{\kappa_0}). \end{aligned} \tag{3.2}$$

with $\mathcal{H} = \{\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0\}$, the hyperparameters for the model parameters. Thus, the generative process for the Bayesian Gaussian mixture model, is rewritten as follows:

$$\begin{aligned} \boldsymbol{\pi}|\boldsymbol{\alpha} &\sim \mathrm{Dir}\left(\alpha_1, \ldots, \alpha_K\right), \\ z_i|\boldsymbol{\pi} &\sim \mathrm{Mult}(1; \pi_1, \ldots, \pi_K), \\ \boldsymbol{\Sigma}_{z_i} &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0), \\ \boldsymbol{\mu}_{z_i}|\boldsymbol{\Sigma}_{z_i} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \tfrac{\boldsymbol{\Sigma}_{z_i}}{\kappa_0}), \\ \mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i} &\sim \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}). \end{aligned} \tag{3.3}$$

Figure 3.2 shows the probabilistic graphical model for the finite Bayesian multivariate GMM.

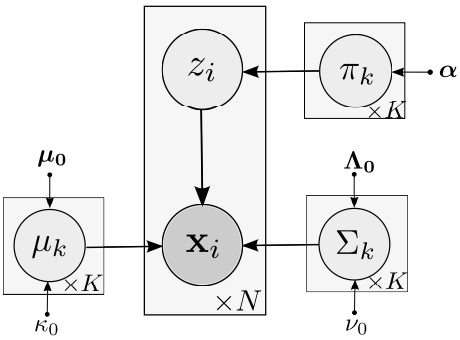


Figure 3.2: *Probabilistic graphical model for the finite Bayesian Gaussian mixture model.*

A detailed description of these densities is given in Gelman et al. (2003). The hyperparameters ν_0 and $\boldsymbol{\Lambda}_0$ describe the degrees of freedom and the scale matrix for the for the inverse-Wishart distribution on $\boldsymbol{\Sigma}$. The remaining hyperparameters are the prior mean, μ_0, and the number of prior measurements, κ_0, on the $\boldsymbol{\Sigma}$ scale. Generally these assumptions are given a priori by the user and are not learned from the data. However, there exists in literature hierarchical Bayesian mixture models (see Richardson and Green (1997); Stephens (1997)) which infer the hyperparameters from the data, making the models more flexible and adaptive for a larger applications variation.

In the next section, we investigate the Bayesian formulation of the parsimonious GMMs, previously described in a ML estimation framework.

3.4 Bayesian parsimonious GMMs

As for the finite Gaussian mixture model, it was natural to derive parsimonious models from the Bayesian GMM, by parametrising the covariance matrix. Fraley and Raftery (2007a, 2005) introduced a Bayesian method by giving prior over the mean vector and the constrained covariance matrix. The authors also discussed the parsimonious Gaussian mixture models extension with the eigenvalue decomposition of the group covariance matrix, $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k$, that was proposed by Banfield and Raftery (1993) and has lead to fourteen models as in Celeux and Govaert (1995). As given in Table 2.2, 14 different flexible Bayesian models were proposed, allowing to vary the volume, orientation and shape of the cluster. Fraley and Raftery (2007a, 2005) provided the priors needed for each of the model parameters, in particular the volume λ, the orientation matrix $\mathbf{D}$ and the shape matrix $\mathbf{A}$. Table 3.5 outlines 14 possible parsimonious Gaussian mixture models, and their respective prior distribution.

Model	Name	Prior	Applied to
$\lambda\mathbf{I}$	EII	$\mathcal{IG}$	λ
$\lambda_k\mathbf{I}$	VII	$\mathcal{IG}$	λ_k
$\lambda\mathbf{A}$	EEI	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
$\lambda_k\mathbf{A}$	VEI	$\mathcal{IG}$ and $\mathcal{IG}$	λ_k and each diagonal element of $\mathbf{A}$
$\lambda\mathbf{A}_k$	EVI	$\mathcal{IG}$ and $\mathcal{IG}$	λ and each diagonal element of $\mathbf{A}$
$\lambda_k\mathbf{A}_k$	VVI	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}_k$
$\lambda\mathbf{DAD}^T$	EEE	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	VEE	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$
$\lambda\mathbf{DA}_k\mathbf{D}^T$	EVE	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}_k$
$\lambda_k\mathbf{DA}_k\mathbf{D}^T$	VVE	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}_k$
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	EEV	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	VEV	$\mathcal{IG}$ and $\mathcal{IW}$	each diagonal element of $\lambda_k\mathbf{A}$ and $\mathbf{D}_k$
$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	EVV	$\mathcal{IG}$ and $\mathcal{IW}$	λ and $\mathbf{\Sigma}_k = \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	VVV	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Table 3.1: *Parsimonious Gaussian Mixture Models via eigenvalue decomposition with the prior associated to each model. Note that $\mathcal{I}$ denotes an inverse distribution, $\mathcal{G}$ denotes a Gamma distribution and $\mathcal{W}$ denotes a Wishart distribution*

3.5 Bayesian inference of the finite mixture model

The Bayesian formulation for mixtures inference is based on estimation of the posterior distributions of the unknown mixture parameters $\boldsymbol{\theta}$, giving the

observed data $\mathbf{X}$ and the prior parameter distribution $p(\boldsymbol{\theta})$. The posterior distribution of the parameters are calculated by Bayes' rule:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{3.4}$$

where the posterior $p(\boldsymbol{\theta}|\mathbf{X})$ is computed by the fraction of the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ penalized by the prior $p(\boldsymbol{\theta})$, and the evidence ($\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})d\boldsymbol{\theta}$). The Bayesian mixture estimation maximizes the posterior (3.4). This is the Maximum A Posteriori (MAP) estimation framework.

The MAP estimation for the Bayesian Gaussian mixture can still be performed, in some situations, by Expectation-Maximization (EM) as in Fraley and Raftery (2007a, 2005); Ormoneit and Tresp (1998); Snoussi and Mohammad-Djafari (2000, 2005). However, the common estimation approach in the case of Bayesian mixtures is Bayesian sampling such as Markov Chain Monte Carlo (MCMC), namely Gibbs sampling (Bensmail et al., 1997; Diebolt and Robert, 1994; Robert, 1994; Stephens, 1997) when the number of mixture components K is known, or by reversible jump MCMC introduced by Green (1995) as in Richardson and Green (1997); Stephens (1997). The flexible eigenvalue decomposition of the group covariance matrix described previously was also exploited in Bayesian parsimonious model-based clustering by Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995) where authors used a Gibbs sampler for the model inference.

3.5.1 Maximum a posteriori (MAP) estimation for mixtures

The Maximum A Posteriori (MAP) estimation framework seeks to estimate the parameters by maximizing the posterior $p(\boldsymbol{\theta}|\mathbf{X})$. Let's denote this posterior distribution function by:

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

then the MAP estimator framework can be summarized as follows:

$$\begin{aligned}
\boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} \mathrm{MAP}(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) \\
&= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})
\end{aligned}$$

One can see, that in Equation (3.4), the denominator, namely the evidence, that is, $\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})d\boldsymbol{\theta}$, is dropped. This is due to the fact that it doesn't depends directly on the parameters $\boldsymbol{\theta}$ on which the maximization is done. Because of numerical computation reasons, the MAP estimator is computed

by maximizing the following logarithm of the posterior parameter distribution:

$$\begin{aligned}\boldsymbol{\theta}_{MAP} &= \arg\max_{\boldsymbol{\theta}} \text{log-MAP}(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \left(\log p(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}|\mathbf{X})\right), \end{aligned} \tag{3.5}$$

where $\log p(\boldsymbol{\theta}|\mathbf{X})$ corresponds to the log-likelihood.

3.5.2 Bayesian inference of the GMMs

For the Bayesian Gaussian mixture model, the MAP estimator framework is then given by the following:

$$\arg\max_{\boldsymbol{\theta}}(\log p(\boldsymbol{\theta}) + \sum_{i=1}^{n}\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \tag{3.6}$$

where the $p(\boldsymbol{\theta})$ is the prior distribution of the model parameters:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\theta}_k), \;\; \boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{3.7}$$

A common choice for the GMM is to assume conjugate priors, that is, a Dirichlet distribution for the mixing proportions $\boldsymbol{\pi}$ (Ormoneit and Tresp, 1998; Richardson and Green, 1997), and a multivariate normal inverse-Wishart prior ($\mathcal{NIW}$) distribution for the Gaussian mixture parameters (Fraley and Raftery, 2007a, 2005; Snoussi and Mohammad-Djafari, 2000, 2005). Thus,

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_0, \kappa_0) p(\boldsymbol{\Sigma}_k|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_0, \nu) \\ &= \text{Dir}(\alpha_1, \ldots, \alpha_K) \prod_{k=1}^{K} \mathcal{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k|\boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Lambda}_0, \nu) \end{aligned} \tag{3.8}$$

This work investigates two approaches for estimation the model parameters in the MAP framework: via the Bayesian Expectation-Maximization algorithm and via the Markov Chain Monte Carlo simulation algorithms.

3.5.3 MAP estimation via the EM algorithm

The Expectation-Maximization algorithm can still be performed for Maximum A Posteriori estimation (MAP) of the Bayesian mixture as in Fraley and Raftery (2007a). Consider the Bayesian Gaussian mixture model discussed previously (3.3). For the Bayesian GMM, the E-step is still the same

as for the ML framework. However, the M-step, depends directly on the penalization term added to the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$. Thus, the M-step for MAP estimation framework updates the mixture parameters by maximizing the following penalized $\mathcal{Q}$ function:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \left[\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) + \log p(\boldsymbol{\theta}^{(t)}) \right] \tag{3.9}$$

This provides the following estimate for the mixture parameters, considered for the M-step (Fraley and Raftery, 2007a, 2005). First, the mixture proportions are updated according to the following:

$$\hat{\pi}_k^{(t+1)} = \frac{n_k^{(t)} + \alpha_k - 1}{n + 1 - K}, \tag{3.10}$$

with n number of observations in data $\mathbf{X}$, $n_k^{(t)}$ the expected number of observations that belongs to the kth component (Equation (2.16)), and K the number of components in the mixture. The mean vector should be updated by it's posterior as follows:

$$\hat{\boldsymbol{\mu}}_k^{(t+1)} = \frac{n_k^{(t)} \bar{\mathbf{x}}_k^{(t)} + \kappa_0 \boldsymbol{\mu}_0}{n_k^{(t)} + \kappa_0}, \tag{3.11}$$

where $\bar{\mathbf{x}}_k^{(t)}$ represents the mean of the data associated to class k, given by the following:

$$\bar{\mathbf{x}}_k^{(t)} = \sum_{i=1}^{n} \frac{\tau_{ik}^{(t)} \mathbf{x}_i}{n_k^{(t)}}.$$

Finally the covariance matrix updated to it's posterior as follows:

$$\hat{\boldsymbol{\Sigma}}_k^{(t+1)} = \frac{\boldsymbol{\Lambda}_0 + \mathbf{W}_k^{(t)} + \frac{\kappa_0 n_k^{(t)}}{n_k^{(t)} + \kappa_0} (\bar{\mathbf{x}}_k^{(t)} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k^{(t)} - \boldsymbol{\mu}_0)^T}{\nu + n_k^{(t)} + d + 2}, \tag{3.12}$$

Recall, $\mathbf{W}_k^{(t)}$ is the scattering matrix of a cluster k given by Equation (2.17). The Bayesian Expectation-Maximization algorithm for the finite mixture model is outlined in the Pseudo-code 3. For a detailed information on the derivation of the EM algorithm in the MAP framework we refer to Fraley and Raftery (2007a, 2005); Ormoneit and Tresp (1998); Snoussi and Mohammad-Djafari (2000, 2005).

3.5.4 Bayesian inference of the parsimonious GMMs via the EM algorithm

As for the MLE framework, where Celeux and Govaert (1995) discussed the EM algorithm for the parsimonious GMMs, it was natural to extend

Algorithm 3 MAP estimation for Gaussian Mixture Models via EM

Inputs: Data set $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, # of mixture components K

Fix: the threshold $\epsilon > 0$, iteration $t \leftarrow 0$ and log-MAP $\leftarrow -\infty$
Initialize $\boldsymbol{\theta}^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \ldots, \boldsymbol{\Sigma}_K^{(0)})$
Initialize the hyperparameters $(\boldsymbol{\alpha}, \boldsymbol{\mu}_0, \kappa_0, \boldsymbol{\Lambda}_0, \nu_0)$.
while increment in log-MAP $> \epsilon$ **do**
  I. E-Step
  for $k \leftarrow 1$ **to** K **do**
    Compute $\tau_{ik}^{(t)}$ $\forall i = 1, \ldots, n$ using Equation (2.11).
  end for
  Compute log-MAP($\boldsymbol{\theta}$) using Equation (3.6).
  II. M-Step
  for $k \leftarrow 1$ **to** K **do**
    Compute $\pi_k^{(t+1)}$ using Equation (3.10).
    Compute $\boldsymbol{\mu}_k^{(t+1)}$ using Equation (3.11).
    Compute $\boldsymbol{\Sigma}_k^{(t+1)}$ using Equation (3.12).
  end for
  $t \leftarrow t + 1$
end while

Outputs: The Gaussian model parameter vector $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t)}$ and the fuzzy partition of the data $\hat{\tau}_{ik} = \tau_{ik}^{(t)}$

the MAP framework estimation via the EM algorithm for the parsimonious GMMs, thus avoiding singularities and degeneracies of the MLE approaches and simultaneously reduce the number of components to estimate. The Maximum A Posteriori (MAP) estimation approach via the EM algorithm presented by Fraley and Raftery (2007a, 2005), discuss the univariate GMMs, as well as the multivariate parsimonious GMMs. The models in Fraley and Raftery (2007a, 2005), are integrated in the existing MCLUST software Fraley and Raftery (1998b, 2007b), which gives the possibility of learning the Bayesian GMMs with the EM algorithm, by taking the eigenvalue parametrization of the covariance matrix $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$. Thus, we implemented the MAP estimation via the EM algorithm for the Parsimonious GMMs. Conjugate prior distributions for the model parameters are used (see for instance Fraley and Raftery (2007a, 2005); Gelman et al. (2003); Ormoneit and Tresp (1998); Snoussi and Mohammad-Djafari (2000, 2005)). The used prior distributions for the decomposed covariance matrix parameters are provided later in Table 3.5.

As the prior distribution does not influence the E-step of the EM algorithm, this step proceeds exactly in the same way as for the MAP framework for the full-GMM model, outlined by Pseudo-code 3. However, the M-step of

the Bayesian EM algorithm varies according to the chosen parametrization of the covariance matrix.

In the M-step of the MAP estimation via EM for parsimonious Bayesian GMMs, the mixture proportions updates are given by Equation (3.10) and the mean vector updates are given by Equation (3.11). However, the M-step, for the covariance matrix, depends on the restricted form of this one. For instance, suppose $\boldsymbol{\Sigma}_k = \lambda\mathbf{I}$, when the spherical covariance matrix with equal volumes is used. In this case, in order to estimate the covariance matrix, the M-step updates only the cluster volume parameter λ. Fraley and Raftery (2007a, 2005) introduces two spherical model, two diagonal model and one general models of the parsimonious multivariate GMMs, that can be easily computed in the MAP framework estimation via the EM. We summarize these models in Table 3.2.

Model		MAP update of $\boldsymbol{\Sigma}_k$
$\lambda\mathbf{I}$	Com-Sphe-GMM	$\frac{\varsigma_0^2+\sum_{k=1}^{K}\mathrm{tr}[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k]}{\nu_0+(n+K)d+2}$
$\lambda_k\mathbf{I}$	Sphe-GMM	$\frac{\varsigma_0^2+\mathrm{tr}[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k]}{\nu_p+(n_k+1)d+2}$
$\lambda\mathbf{A}$	Com-Diag-GMM	$\frac{\mathrm{diag}(\varsigma_0^2\mathbf{I}+\sum_{k=1}^{K}[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k])}{\nu_0+n+K+2}$
$\lambda_k\mathbf{A}_k$	Diag-GMM	$\frac{\mathrm{diag}(\varsigma_0^2\mathbf{I}+[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k])}{\nu_0+n_k+3}$
$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T$	Com-GMM	$\frac{\boldsymbol{\Lambda}_0+\sum_{k=1}^{K}[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k]}{\nu_0+n+d+K+1}$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Full-GMM	$\frac{\boldsymbol{\Lambda}_0+[\frac{\kappa_0 n_k}{n_k+\kappa_0}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T+\mathbf{W}_k]}{\nu_p+n_k+d+2}$

Table 3.2: *M-step estimation for the covariances of multivariate mixture models under the Normal inverse Gamma conjugate prior for the spherical models ($\lambda\mathbf{I}$, $\lambda_k\mathbf{I}$) and the diagonal models ($\lambda\mathbf{A}$, $\lambda_k\mathbf{A}_k$), and Normal inverse Wishart conjugate priors for the general models ($\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T$, $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$).*

The hyperparameters are usually chosen a priori by the user and not learned from the data. This is also the case in the study of Fraley and Raftery (2007a, 2005). Thus, choosing good values for hyperparameters that are adaptive to a particular data is one important issue in this Bayesian learning framework. The following choices for hyperparameters of the multivariate Bayesian GMMs were found effective in the experimentations of Fraley and Raftery (2007a, 2005):

- $\boldsymbol{\mu}_0$ is considered to be equal to the mean of the data.
- κ_0 is considered to be equal to 0.01. The posterior of the mean can be viewed as adding κ_0 observations to the $\boldsymbol{\mu}_0$ value of each group of data.

- ν_0 which can be interpreted as the degrees of freedom of the model, is chosen to be the minimum integer value for the degrees of freedom, that is equal to $\nu_p = d + 2$ (Schafer, 1997).
- ς_0^2 that we need to calculate in the case of spherical covariances models are assumed to be equal to $\varsigma_p^2 = \frac{\text{sum}(\text{diag}(\text{cov}(\mathbf{X})))/d}{K^{2/d}}$.
- $\mathbf{\Lambda}_0$, used for the general models, is computed by $\mathbf{\Lambda}_0 = \frac{\text{cov}(\mathbf{X})}{K^{2/d}}$.

When the posterior distributions can not be analytically computed, Markov Chain Monte Carlo (MCMC) methods can be used. Next, we investigate the Bayesian inference via the MCMC methods.

3.5.5 Markov Chain Mote Carlo (MCMC) inference

The common estimation approach in the case the Bayesian mixture models described above, is the one using Bayesian sampling such as Markov Chain simulations, also called in literature as Markov Chain Monte Carlo (MCMC) sampling techniques (Bensmail and Meulman, 2003; Bensmail et al., 1997; Diebolt and Robert, 1994; Escobar and West, 1994; Geyer, 1991; Gilks et al., 1996; Neal, 1993; Richardson and Green, 1997; Robert, 1994; Stephens, 1997).

The Markov chain is known as a sequence of random variables, $\boldsymbol{\theta}^{(t)}$, such that $t \geq 1$, where each of tth variable distribution depends only on the previous $t-1$ variable distribution. So the basic idea of the Markov chain Monte Carlo inference methods is to obtain the ergodic Markov chain by drawing sequentially the mixture parameters $\boldsymbol{\theta}$ from an approximate distributions $p(\boldsymbol{\theta}^{(t)}|\mathbf{X})$, to better approximate the expected posterior distribution $E[p(\boldsymbol{\theta}|\mathbf{X})]$.

$$\begin{aligned} E[p(\boldsymbol{\theta}|\mathbf{X})] &= \int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &\approx \frac{1}{n_s}\sum_{t=1}^{n_s} p(\boldsymbol{\theta}^{(t)}|\mathbf{X}) \end{aligned} \tag{3.13}$$

The starting point $\boldsymbol{\theta}^{(0)}$ influences directly the MCMC convergence speed. Also, the approximation of the posterior distribution, given in Equation (3.13), becomes more precise when the number of samples n_s, goes to infinity (Meyn and Tweedie, 1993), so a big number of samples n_s provide a better posterior approximation. The idea of using such MCMC methods, dates back to early Physics literature Metropolis et al. (1953) when the computational power was not even available. This provides a generic sampling method, namely the Metropolis-Hashing algorithm Hastings (1970); Metropolis et al. (1953).

A widely used method for MCMC sampling is the Gibbs sampling. This work investigates the Gibbs sampling algorithm, for the Bayesian inference

of the Gaussian mixture model. In particular, the inference of the Bayesian parsimonious GMMs via the Gibbs sampling is presented and discussed. The Gibbs sampling takes it's name referencing to the name of Gibbs random fields used by Geman and Geman (1984), that was proposed in a framework of Bayesian image restoration. A very close form to it was also introduced by Tanner and Wong (1987) under the name of data augmentation for missing data problems, and shown in Gelfand and Smith (1990). For more details on Gibbs sampling we also refer to Casella and George (1992); Diebolt and Robert (1994); Gelfand et al. (1990); Gilks et al. (1996); Marin and Robert (2007); Robert (1994).

Suppose a hierarchical structure of the model where the posterior can be given by:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \int p(\boldsymbol{\theta}|\mathbf{X}, \mathcal{H})p(\mathcal{H}|\mathbf{X})d\mathcal{H} \tag{3.14}$$

where $\mathcal{H}$ are the hyperparameters of the model parameters $\boldsymbol{\theta}$. The idea of Gibbs sampling is then to simulate from the joint distribution $p(\boldsymbol{\theta}|\mathbf{X}, \mathcal{H})p(\mathcal{H}|\mathbf{X})$, to approximate better the posterior $p(\boldsymbol{\theta}|\mathbf{X})$. Assuming that these distributions are known, the parameters $\boldsymbol{\theta}$ and hyperparameters $\mathcal{H}$, shall be drawn respectively by the $p(\boldsymbol{\theta}|\mathbf{X}, \mathcal{H})$ and $p(\mathcal{H}|\mathbf{X})$. However, more generally the hyperparameters $\mathcal{H}$ are supposed to be known and given a priori by the user, so that only the parameters $\boldsymbol{\theta}$ are sampled.

The general Gibbs sampling algorithm for the mixture models, therefore simulates the joint distribution $p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ from the full conditional distribution $p(\boldsymbol{\theta}_k|\{\boldsymbol{\theta}\}_{\setminus\boldsymbol{\theta}_k}, \mathbf{X})$ as outlined in Pseudo-code 4.

Algorithm 4 Gibbs sampling for mixture models

Input: The data set $\mathbf{X} = (\mathbf{x}_1, \ldots \mathbf{x}_n)$, # of mixture components K and # of samples n_s.
Initialize the model parameters $\boldsymbol{\theta}^{(0)}$.
for $t = 1$ **to** n_s **do**
  for $k = 1$ **to** K **do**
    Sample $\boldsymbol{\theta}_k^{(t)}$ from the posterior distribution $p(\boldsymbol{\theta}_k|\{\boldsymbol{\theta}\}_{\setminus\boldsymbol{\theta}_k}^{(t-1)}, \mathbf{X})$
  end for
end for
Outputs: The Markov chain parameters vector of the mixture $\hat{\boldsymbol{\Theta}} = \boldsymbol{\theta}^{(t)}, \ \forall t = 1, \ldots, n_s$.

One debate for the MCMC methods (e.g. Gibbs sampling), is the convergence. The speed of the convergence depends directly on the initialization step. Also having a good initialization of the model parameters tackle a smaller burn-in period. The initialization step, that computes the initial parameter vector, can be done by:

- running itself the Gibbs sampling, this can be investigated by running many short chains as in Gelfand and Smith (1990) or few long chains as in Gelman and Rubin (1992),
- random initialization, this usually needs one vary long chains as in (Geyer, 1992) and a long burn-in period,
- running other clustering algorithms like K-means initialization (MacQueen, 1967), that is the case of this work.

Later in our experiments we see that, usually 10-20 chains with 2000 Gibbs samples is sufficient. Also, because the first simulations depend directly on the initialization $\boldsymbol{\theta}^{(0)}$, normally they are not fitting very well the mixture model. Therefore, a burn-in period can be considered, that generally takes 10% for the number of samples. Also, in practice it is usually proposed to run multiple Gibbs samplings where different initialization for the model parameters $\boldsymbol{\theta}^{(0)}$ are proposed.

3.5.6 Bayesian inference of GMMs via Gibbs sampling

Here we investigate the Gibbs sampling for the multivariate Gaussian mixture model that we examine in detail for this work. Suppose the Bayesian GMM given in Equation (3.3), where the mixture parameters are $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ with $\boldsymbol{\theta}_k = \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, $\forall k = 1, \ldots, K$. The Gibbs sampler for GMMs is the following Pseudo-code 5. One can see, in Pseudo-code 5, that the labels z_i and the mixture parameters $\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ are sampled respectively by Mult(.), Dir(.), $\mathcal{N}(.)$ and $\mathcal{IW}(.)$, that are the Multinomial, Dirichlet, Normal and inverse Wishart distributions. Their detailed mathematical computation can be found in Appendix (B). Also, $\{\boldsymbol{\mu}_n, \kappa_n, \nu_n, \boldsymbol{\Lambda}_n\}$ are the respective posterior for the hyperparameters $\{\boldsymbol{\mu}_0, \kappa_0, \nu_0, \Lambda_0\}$. As proposed by Gelman et al. (2003), the computation of the hyperparameters posterior is then given by:

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{n_k \bar{\mathbf{x}}_k + \kappa_0 \boldsymbol{\mu}_0}{n_k + \kappa_0} \\
\kappa_n &= \kappa_0 + n_k \\
\nu_n &= \nu_0 + n_k \\
\boldsymbol{\Lambda}_n &= \boldsymbol{\Lambda}_0 + \mathbf{W}_k + \frac{n_k \kappa_n}{n_k + \kappa_n} (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T
\end{aligned} \tag{3.15}$$

Note that, the parameter vector is obtained by averaging the Gibbs samples after removing a burn-in period.

3.5.7 Illustration: Bayesian inference of the GMM via Gibbs sampling

We implement the Gibbs sampling approach and show it's effectiveness for estimating the Gaussian mixture model. First, we considered a two-class

Algorithm 5 Gibbs sampling for Gaussian mixture models

Input: The data set $\mathbf{X} = (\mathbf{x}_1, \dots \mathbf{x}_n)$, # of mixture components K, # of samples n_s.

Initialize: The hyperparameter $\mathcal{H}^{(0)} = (\boldsymbol{\alpha}^{(0)}, \boldsymbol{\mu}_0^{(0)}, \kappa_0^{(0)}, \boldsymbol{\Lambda}_0^{(0)}, \nu_0^{(0)})$, the mixture probabilities $\boldsymbol{\pi}^{(0)}$, and the component parameters $\boldsymbol{\theta}_k^{(0)} = \{\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}\}$

for $t = 1$ **to** n_s **do**

  for $k = 1$ **to** K **do**

    1. Sample the labels $z_i^{(t)} | \tau_{ik}^{(t)}, \boldsymbol{\pi}_k^{(t-1)}, \boldsymbol{\theta}_k^{(t-1)} \sim \text{Mult}(1, \tau_{i1}^{(t)}, \dots, \tau_{iK}^{(t)})$ conditional on the posterior probabilities $\tau_{ik}^{(t)} = \frac{\pi_k^{(t-1)} \mathcal{N}_k(\mathbf{x}_i | \boldsymbol{\theta}_k^{(t-1)})}{\sum_{k=1}^{K} \pi_k^{(t-1)} \mathcal{N}_k(\mathbf{x}_i | \boldsymbol{\theta}_k^{(t-1)})}$.

  end for

end for

2. Sample the mixture probabilities according to the posterior distribution $\boldsymbol{\pi}^{(t)} | \tau_{ik}^{(t)}, \boldsymbol{\mu}_k^{(t-1)}, \boldsymbol{\Sigma}_k^{(t-1)}, \mathbf{X} \sim \text{Dir}(\alpha_1 + n_1, \dots, \alpha_K + n_K)$.

for $t = 1$ **to** n_s **do**

  for $k = 1$ **to** K **do**

    3. Sample the mean vector $\boldsymbol{\mu}_k^{(t)}$ according to the posterior distribution $\boldsymbol{\mu}_k^{(t)} | \tau_{ik}^{(t)}, \boldsymbol{\pi}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t-1)}, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_k / \kappa_n)$.

    4. Sample the covariance matrix $\boldsymbol{\Sigma}_k^{(t)}$ according to the posterior distribution $\boldsymbol{\Sigma}_k^{(t)} | \tau_{ik}^{(t)}, \boldsymbol{\pi}_k^{(t)}, \boldsymbol{\mu}_k^{(t)}, \mathbf{X} \sim \mathcal{IW}(\nu_n, \boldsymbol{\Lambda}_n)$.

  end for

end for

Outputs: The parameters vector chain of the mixture $\hat{\boldsymbol{\Theta}} = \{\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}, \ \forall t = 1, \dots, n_s$.

situation identical to the one in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995) where parametric parsimonious mixture approach (see Subsection 3.5.8) is proposed. The data consist in a sample of $n = 200$ observations from a two-component Gaussian mixture in $\mathbb{R}^2$ with the following parameters: equal mixture proportions $\pi_1 = \pi_2 = 0.5$, the mean vectors $\boldsymbol{\mu}_1 = (8, 8)^T$ and $\boldsymbol{\mu}_2 = (2, 2)^T$, and two spherical covariances with different volumes $\boldsymbol{\Sigma}_1 = 4\,\mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \mathbf{I}_2$. An illustration of this dataset can be seen in the Figure 3.3. For this experiment, we sampled 2000 Gibbs samples, ten times, with 10% burn-in, for the finite Bayesian Gaussian mixture model. The obtained partition is given in Figure 3.4. The estimated model parameter values are $\hat{\boldsymbol{\pi}} = (0.5285, 0.4715)^T$ $\hat{\boldsymbol{\mu}}_1 = (7.9631, 8.0156)^T$ and $\hat{\boldsymbol{\mu}}_2 = (1.8890, 2.0389)^T$, and $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 4.9511 & -0.1054 \\ -0.1054 & 3.3794 \end{pmatrix}$, and $\boldsymbol{\Sigma}_2 = \begin{pmatrix} 1.2585 & 0.2583 \\ 0.2583 & 1.2250 \end{pmatrix}$. The estimates are close to the actual parameters.

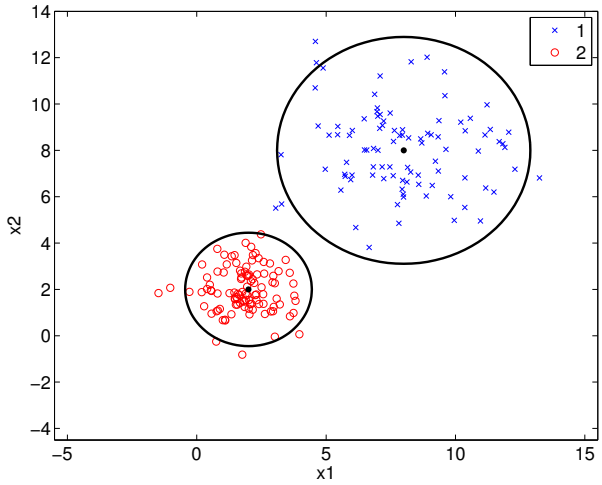


Figure 3.3: *A simulated dataset from a mixture model in $\mathbb{R}^2$ two component Gaussian.*

In order to evaluate our clustering, we use the error rate that is the error computed between the true (simulated) and the estimated labels of the data. On the other hand, we evaluate our clustering with the Rand index (Rand, 1971) values. For a more variety of the different clustering indexes and their mathematical computation we refer to Desgraupes (2013). In Figure 3.4, one can see the error rate (on middle) and respective the Rand index (on right) values are computed for each sample of the Gibbs method. Highlight the fact that the best obtained value for the error rate is equal to zero, meaning that all the estimated labels are equivalent to the true labels, while the best value for the Rand index is equal to one.

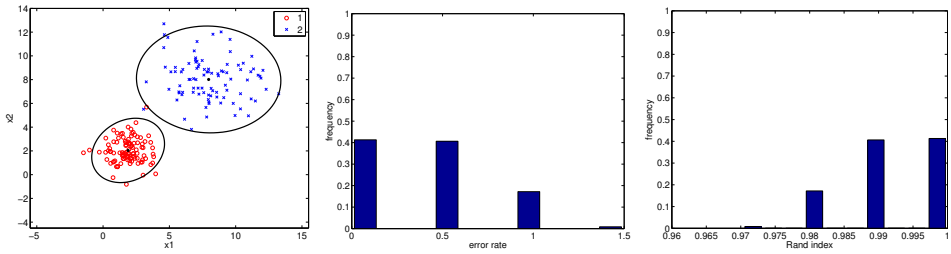


Figure 3.4: *The Gibbs sampling for the Full-GMM model of the dataset shown in Figure 3.3, with the estimated partition (left), the obtained error rate (middle) and the Rand Index (right).*

In order to compare with future results, obtained by the Parsimonious GMMs, discussed in Subsection 3.5.8, we give, in Table 3.3, the resulted values for the marginal likelihood (ML), log-MAP, Rand index (RI), error rate (ER) values, the number of parameters to estimate and the Gibbs sampler

time processing (in seconds). Note that the marginal likelihood is mostly needed for the Bayes factor computation, that offers a Bayesian comparison and selection of the models. We discuss this in detail in Subsection 3.5.9.

ML	log-MAP	RI	ER	# parameters	Cpu time (s)
-861.6041	-855.38	1	0	11	145.72

Table 3.3: *The obtained marginal likelihood (ML), log-MAP, Rand index (RI), error rate (ER) values, the number of parameters to estimate and the time processing (in seconds) for the Gibbs sampling for GMM for the two class simulated dataset.*

We also applied the Gibbs sampler with two components Full-GMM to the Old Faithful Geyser and Iris dataset. The obtained results are given in Figure 3.5.

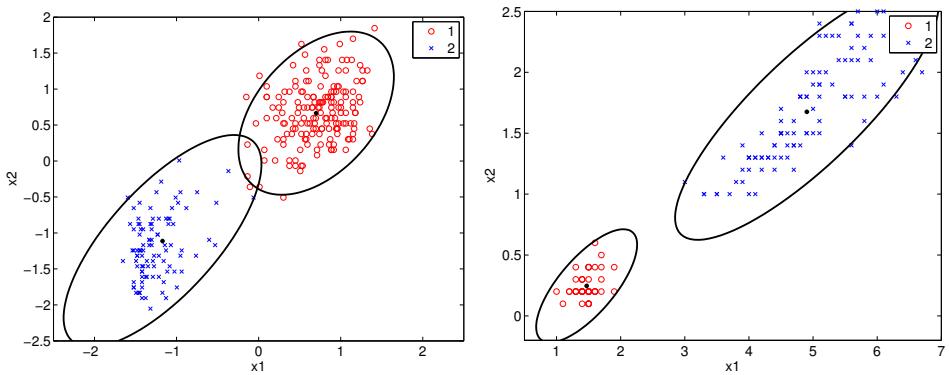


Figure 3.5: *Gibbs sampling partitions and model estimates for a two-component full-GMM model obtained for the Old Faithful Geyser dataset (left) and Iris dataset (right).*

A numerical computation for the Old Faithful Geyser, and Iris dataset obtained by learning the two component Full-GMM with the Gibbs sampling approach, is given by the marginal likelihood (ML), log-MAP, Rand index (RI), error rate (ER) values, the number of parameters to estimate and the Gibbs sampler time processing (in seconds). This is provided in Table 3.4.

Data et	ML	log-MAP	# parameters	Cpu time (s)
Old Faithful Geyser	-428.60	-409.83	11	146.46
Iris	-272.88	-223.38	29	68.52

Table 3.4: *The obtained marginal likelihood (ML), log-MAP, the number of parameters to estimate and the time processing (in seconds) for the Gibbs sampling GMM on the Old Faithful Geyser and Iris dataset.*

Naturally, the Gibbs sampling for Parsimonious GMMs was investigated, and we study it in the next subsection.

3.5.8 Bayesian inference of parsimonious GMMs via Gibbs sampling

As outlined in Bensmail et al. (1997), the approach of Banfield and Raftery (1993) that infers the parsimonious mixture with EM algorithm has some limitations, for example: no assessment of the uncertainty about the classification, as it gives only point estimation, the shape matrix has to be specified by the user, prior group probabilities are assumed to be equal, etc. Thus, Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995) proposed a new Bayesian approach which overcomes these difficulties. This approach consists in exact Bayesian inference via Gibbs sampling and the calculation of Bayes Factors that is used for simultaneously choosing the models and the number of groups. The computation of the Bayes Factor is based on the Laplace-Metropolis estimator (Lewis and Raftery, 1994; Raftery, 1996), where the marginal likelihood is computed via the posterior simulation output.

Consider the Bayesian inference for the multivariate parsimonious Gaussian mixture model, with the eigenvalue decomposition of the covariance matrix. Recall that the MCMC approaches provide methods for estimating the model consisting of: the partitions $\mathbf{z} = \{z_1, \ldots, z_n\}$ and the mixture parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ where for each group k we have the mean vector and the covariance matrix: $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995) used conjugate priors for the model parameters $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ as in Diebolt and Robert (1994); Tanner and Wong (1987), where the prior distributions over the mixture proportions $\boldsymbol{\pi}$ is a Dirichlet distribution, $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, with $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_k\}$ and the prior distribution for the mean vector, conditional on the covariance matrix is a multivariate normal distribution, $\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k/\kappa_0)$. The prior for the covariance matrix $\boldsymbol{\Sigma}_k$ depends on the selected parsimonious GMM. Therefore, the simulation step for this parameter varies according to the given priors. Table 3.5 gives the prior for the different parsimonious GMMs used in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995), where eigenvalue decomposition for the covariance matrix is considered.

Also, the model selection problem was considered in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995), where the approximate Bayes Factors from the Gibbs sampler output using the Laplace-Metropolis estimator was used to simultaneously choose the number of groups and the eigenvalue decomposition of the parsimonious GMM. On the other hand, in order to facilitate the task of computing the marginal likelihoods, information criteria can be also used in the Bayesian inference, like MCMC algorithms, to compare performance of the different competitive models (see

Model	Prior	Applied to
$\lambda\mathbf{I}$	$\mathcal{IG}$	λ
$\lambda_k\mathbf{I}$	$\mathcal{IG}$	λ_k
$\lambda\mathbf{DAD}^T$	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\mathbf{\Sigma} = \mathbf{DAD}^T$
$\lambda\mathbf{DA}_k\mathbf{D}^T$	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}_k$
$\lambda_k\mathbf{DA}_k\mathbf{D}^T$	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}_k$
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}$
$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	$\mathcal{IG}$ and $\mathcal{IW}$	λ and $\mathbf{\Sigma}_k = \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Table 3.5: *Bayesian Parsimonious Gaussian mixture models via eigenvalue decomposition with the associated prior as in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995).*

for example Biernacki and Govaert (1998)).

In the next section, we present the model selection and comparison in the Bayesian formulation and investigate it's use for mixture models, including Gaussian mixtures and their parsimonious counterparts.

3.5.9 Bayesian model selection and comparison using Bayes Factors

For the non-Bayesian parametric approach, one important task is the estimation of number of components in the mixture. This issue is also encountered in the Bayesian context, referred to as the Bayesian model selection (Wasserman, 2000). We discussed that for the MAP approach where, the choice of the optimal number of mixture components and the best model structure, can still be performed via modified penalized log-likelihood criteria such as a modified version of BIC as in (Fraley and Raftery, 2007a) computed for the posterior mode. In this section, we discuss a more general Bayesian approach that is the Bayes Factors (Kass and Raftery, 1995).

The problem of the model selection in the finite Bayesian mixture model-based clustering can be tackled by generally using the Bayes Factors (Kass and Raftery, 1995), as in Bensmail et al. (1997); Bensmail (1995). Bayes Factors provide a general way to select and compare the models in (Bayesian) statistical modeling by comparing the marginal likelihood of the models. They have been widely studied in the case of mixture models (Basu and Chib, 2003; Bensmail et al., 1997; Carlin and Chib, 1995; Gelfand and Dey, 1994; Kass and Raftery, 1995; Raftery, 1996).

Suppose that we have two models candidates, M_1 and M_2, the Bayes factor is given by:

$$BF_{12} = \frac{p(\mathbf{X}|M_1)p(M_1)}{p(\mathbf{X}|M_2)p(M_2)}. \tag{3.16}$$

In this work, we assume that the two models have the same prior probability $p(M_1) = p(M_2)$. The Bayes factor (3.16) is thus given by

$$BF_{12} = \frac{p(\mathbf{X}|M_1)}{p(\mathbf{X}|M_2)}, \tag{3.17}$$

which corresponds to the ratio between the marginal likelihoods of the two models M_1 and M_2. It is a summary of the evidence for model M_1 against model M_2 given the data $\mathbf{X}$. Note that, often, for numerical computational reasons, the logarithm of the Bayes Factor is considered:

$$\log BF_{12} = \log p(\mathbf{X}|M_1) - \log p(\mathbf{X}|M_2). \tag{3.18}$$

The marginal likelihood $p(\mathbf{X}|M_m)$ for model M_m, $m \in \{1, 2\}$, also called the integrated likelihood, is given by

$$p(\mathbf{X}|M_m) = \int p(\mathbf{X}|\boldsymbol{\theta}_m, M_m) p(\boldsymbol{\theta}_m|M_m) d\boldsymbol{\theta}_m \tag{3.19}$$

where $p(\mathbf{x}|\boldsymbol{\theta}_m, M_m)$ is the likelihood of model M_m with parameters $\boldsymbol{\theta}_m$ and $p(\boldsymbol{\theta}_k|\mathcal{M}_m)$ is the prior density of the parameters $\boldsymbol{\theta}_m$ of model M_m. As we can see in Equation (3.19), the existence of the integral, makes difficult the analytic calculation of the marginal likelihood. Therefore, several MCMC approximation methods have been proposed to estimate the marginal likelihood. One of the simplest, is by sampling the parameters $\boldsymbol{\theta}$ from the prior distribution and approximating the marginal likelihood as:

$$\hat{p}_{PR}(\mathbf{X}|M_m) = \frac{1}{n_s} \sum_{t=1}^{n_s} p(\mathbf{X}|M_m, \boldsymbol{\theta}_m^{(t)}) \tag{3.20}$$

where n_s is the number of MCMC samples, the model parameters $\boldsymbol{\theta}_m^{(t)}$ are sampled according to the prior distributions. This computation can be seen as the empirical mean of the likelihood values (Hammersley and Handscomb, 1964). However, this is an unstable and inefficient method, that needs a lot of running time (Bensmail, 1995). Therefore, a wide number of alternative methods were proposed to compute the marginal likelihood according to the posterior distribution, instead of the prior distribution (M. and Roberts, 1993; Newton and Raftery, 1994; Rubin, 1987; Tanner and Wong, 1987). The harmonic mean of the likelihood values computes the marginal likelihood (Newton and Raftery, 1994) as follows:

$$\hat{p}_{HM}(\mathbf{X}|M_m) = \left\{ \frac{1}{n_s} \sum_{t=1}^{n_s} p(\mathbf{X}|\boldsymbol{\theta}_m^{(t)})^{-1} \right\}^{-1}. \tag{3.21}$$

This converges practically in a correct value of marginal likelihood $p(\mathbf{X}|M_m)$ as the number of MCMC samples becomes high. However, it can lead to

unstable results. A modification of Equation (3.21), was than proposed, to give more accurate solution for the resulting estimated marginal likelihood (Gelfand and Dey, 1994). The approximation of the marginal likelihood, in this case, is given by

$$\hat{p}_{GD}(\mathbf{X}|M_m) = \frac{1}{n_s}\sum_{t=1}^{n_s}\left\{\frac{p(\boldsymbol{\theta}_m^{(t)}|\mathbf{X})}{p(\mathbf{X}|\boldsymbol{\theta}_m^{(t)})p(\boldsymbol{\theta}_m^{(t)})}\right\}. \tag{3.22}$$

Another estimation of the marginal likelihood with Gibbs sampling from the posterior was proposed by Chib (1995), where he uses directly the Bayes rule, to get the marginal likelihood. The resulting approximation of the marginal likelihood is then given by

$$\hat{p}_{\text{Chib}}(\mathbf{X}|M_m) = \frac{p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m)p(\hat{\boldsymbol{\theta}}_m)}{\prod_{i=1}^{n_s} p(\hat{\boldsymbol{\theta}}_m^{(i)}|\mathbf{x}, \hat{\boldsymbol{\theta}}_m^{(j)}(j<i))}. \tag{3.23}$$

Finally, one more accurate approximation of the marginal likelihood, by estimating consecutively the posterior of the model parameters with the Gibbs sampling, is the Laplace-Metropolis approximation (Lewis and Raftery, 1994; Raftery, 1996). This method shown to give accurate results in (Lewis and Raftery, 1994; Raftery, 1996) and then used as the Bayesian model selection in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995) giving appropriate results for the parsimonious models that we assume in this work, thus we investigated it more in details future in our experimentations. The equation computing the marginal likelihood can be summarized by:

$$\hat{p}_{\text{Laplace}}(\mathbf{X}|M_m) = (2\pi)^{\frac{\nu_m}{2}}|\hat{\mathbf{H}}|^{\frac{1}{2}}p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m) \tag{3.24}$$

where $\hat{\boldsymbol{\theta}}_m$ is the posterior estimation of $\boldsymbol{\theta}_m$ (posterior mode) for model M_m, ν_m is the number of free parameters of the model M_m as given, for example Table 4.1 for the mixture case, and $\hat{\mathbf{H}}$ is minus the inverse Hessian of the function $\log(p(\mathbf{X}|\hat{\boldsymbol{\theta}}_m, M_m)p(\hat{\boldsymbol{\theta}}_m|M_m))$ evaluated at the posterior mode of $\boldsymbol{\theta}_m$, that is $\hat{\boldsymbol{\theta}}_m$. The matrix $\hat{\mathbf{H}}$ is asymptotically equal to the posterior covariance matrix (Lewis and Raftery, 1994), and is computed as the sample covariance matrix of the posterior simulated sample.

Once the estimation of Bayes Factors is obtained, it can be interpreted as described in Table 3.6 as suggested by Jeffreys (1961), see also Kass and Raftery (1995).

Bayes factors are indeed the natural Bayesian criterion for model selection and comparison in the Bayesian framework and for which the criteria such as BIC, AWE, etc represent approximations. The computation of these criteria, namely the information criteria, are more simple and doesn't need the computation of the marginal likelihood.

BF_{12}	$2\log \mathrm{BF}_{12}$	Evidence for model M_1
< 1	< 0	Negative (M_2 is selected)
$1-3$	$0-2$	Not bad
$3-12$	$2-5$	Substantial
$12-150$	$5-10$	Strong
> 150	> 10	Decisive

Table 3.6: *Model comparaion and selection using Bayes factors.*

3.5.10 Experimental study

The parsimonious models Celeux and Govaert (1995), where some of them have been described in the Bayesian framework in Bensmail and Meulman (2003); Bensmail et al. (1997); Bensmail (1995), have been all derived in a Bayesian framework in this thesis and all implemented in MATLAB. In this section, we experiment the Bayesian parsimonious models on simulations in order to assess them in terms of model estimation selection and comparison. We also consider application on a Old Faithful Geyser dataset.

Generally the Bayesian mixture model, which we investigate here, is not a hierarchical model, the hyperparameters being known and given a priori by the user. It is important and a challenging problem to find the best hyperparameters values that fit at best the data. In this experimental study we investigate the influence of changing the hyperparameter values on the final result. This can be seen, somehow, as a model selection problem. The final partitions are also assessed for the Gibbs sampling for the parsimonious GMMs.

Consider the two spherical class dataset presented in subsection (3.5.7), where the true model parameters are $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (8,8)^T$ and $\boldsymbol{\mu}_2 = (2,2)^T$ and two spherical covariance matrices with different volumes: $\boldsymbol{\Sigma}_1 = 4\,\mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \mathbf{I}_2$. We use the implemented Gibbs sampling algorithm for parameter estimation. In order to assess the stability of the models with respect to the values of the hyperparameters, we consider four situations with different hyperparameter values. These are as follows. The hyperparameters ν_0 and $\boldsymbol{\mu}_0$ are assumed to be the same for the four situations and their values are respectively $\nu_0 = d+2 = 4$ (related to the number of degrees of freedom) and $\boldsymbol{\mu}_0$ equals the empirical mean vecotr of the data. We variate the two hyperparameters, κ_0 that controls the prior over the mean and s_0^2 that controls the covariance. The considered four situations are shown in Table 5.12.

The Gibbs sampler is run to sample 2000 Gibbs samples, for each of these models, ten times, with 10% burn-in, for the finite parsimonious Gaussian mixture models. We also vary the number of components in the mixture, from one to five, $K = 1, \ldots, 5$. The best model, that fits at best the data, that includes the best number of components and the best model structure,

Sit.	1	2	3	4
s_0^2	$\max(\text{eig}(\text{cov}(\mathbf{X})))$	$\max(\text{eig}(\text{cov}(\mathbf{X})))$	$4\ \max(\text{eig}(\text{cov}(\mathbf{X})))$	$\max(\text{eig}(\text{cov}(\mathbf{X})))/4$
κ_0	1	5	5	5

Table 3.7: *Four different situations the hyperparameters values.*

is then selected according to the maximum marginal log-likelihood (Bayes Factors). We consider and compare the four following models the spherical, diagonal and general models, which correspond to, respectively, $\lambda\mathbf{I}$, $\lambda_k\mathbf{I}$, $\lambda\mathbf{A}$ and $\lambda_k\mathbf{DAD}^T$.

Figure 3.6 shows the model selection results for the four hyperparameters varying situations and for a number of components varying from one to five, $(K = 1, \ldots, 5)$. One can see that the actual spherical model $\lambda_k\mathbf{I}$ with the three number of components, was selected for the four situations. Another model, that can be considered to be the most competitive one, is the general model with different volumes and the same orientation and shape between the clusters $(\lambda_k\mathbf{DAD}^T)$.

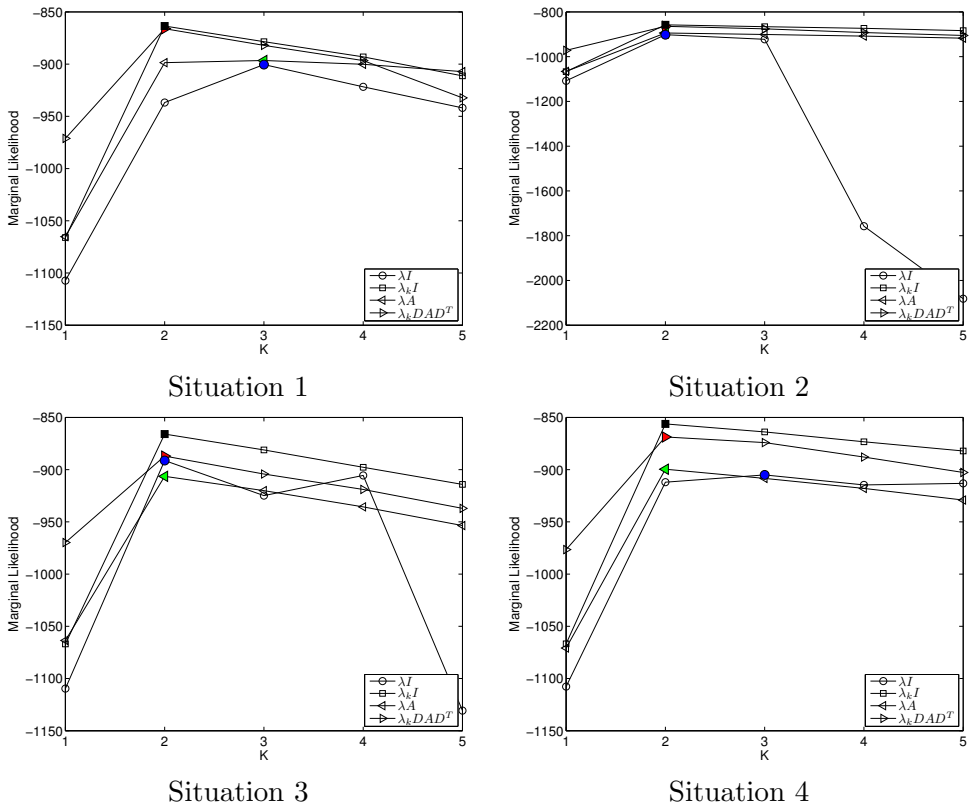


Figure 3.6: *Model selection with marginal log-likelihood for the two component spherical dataset represented in Figure 3.3.*

Table 3.8 shows the obtained marginal log-likelihood values for the four models for the for situations of varying the hyperparameters shown in Table 5.12. One can see that, according to the marginal log-likelihood, for all the situations, the selected model is $\lambda_k\mathbf{I}$, that is the one that corresponds to the actual model, and has the correct number of mixture components (two). Also, the models with the model structure with varying volumes ($\lambda_k\mathbf{I}$ and $\lambda_k\mathbf{DAD}^T$) estimate a good number of clusters for the four situations, meaning a stability over the variation of the hyperparameters.

Model	$\lambda\mathbf{I}$		$\lambda_k\mathbf{I}$		$\lambda\mathbf{A}$		$\lambda_k\mathbf{DAD}^T$	
Sit.	$\hat{K}$	log ML	$\hat{K}$	log ML	$\hat{K}$	log ML	$\hat{K}$	log ML
1	3	-900.4241	2	$\mathbf{-863.5121}$	3	-896.5311	2	-866.0787
2	2	-901.8706	2	$\mathbf{-857.9103}$	2	-894.2924	2	-864.4517
3	2	-891.2702	2	$\mathbf{-865.9100}$	2	-906.4263	2	-887.0174
4	3	-905.0301	2	$\mathbf{-856.2335}$	2	-899.5766	2	-868.6876

Table 3.8: *The marginal log-likelihood values for the finite and infinite parsimonious Gaussian mixture models.*

Additionally, Figure 3.7 shows the obtained partition for the fourth hyperparameter settings of Table 5.12 for different models. One can see different geometrical forms corresponding to the different parsimonious models. On top left the spherical covariance with equal volumes. On top right, the best selected model that also corresponds to the actual model with spherical covariance and different volumes. On bottom left, the diagonal model with equal volume and the same shape, is represented. Finally the general model with different volume but the same shape and orientation of the covariance matrix structure can be observed on the bottom right of the figure.

In addition to the simulated data experiment discussed previously, we also apply the implemented Gibbs sampling for the parsimonious GMMs on the well known dataset, the Old Faithful Geyser data, shown in Figure 2.7. The hyper-parameters for the treated parsimonious GMMs are set as follows: $\kappa_0 = 5$, $\nu = d + 2$, $\mathbf{\Lambda}_0$ is equal to the covariance of the data and s_0^2 is the maximum eigenvalue of the covariance of the data. We vary the number of clusters K from 1 to 10 for model selection. Five models, with the following eigenvalue covariance decomposition, are studied in this experiment: $\lambda_k\mathbf{I}$, $\lambda_k\mathbf{A}$, $\lambda\mathbf{DAD}^T$, $\lambda_k\mathbf{DAD}^T$ and the Full-GMM $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$.

First, Figure 3.8 shows the model selection results by using the marginal log-likelihood given in Equation (3.24). One can see that, except the Full-GMM that overestimates the number of components ($\hat{K} = 5$), the other models select the number of components ($\hat{K} = 2$). The best model that is the one with the covariance decomposition $\lambda_k\mathbf{DAD}^T$ (a different volumes but equal orientations and shapes for the components).

As previously mentioned, the computation of the marginal likelihood can be simplified by computing approximations for Bayes Factors, namely infor-

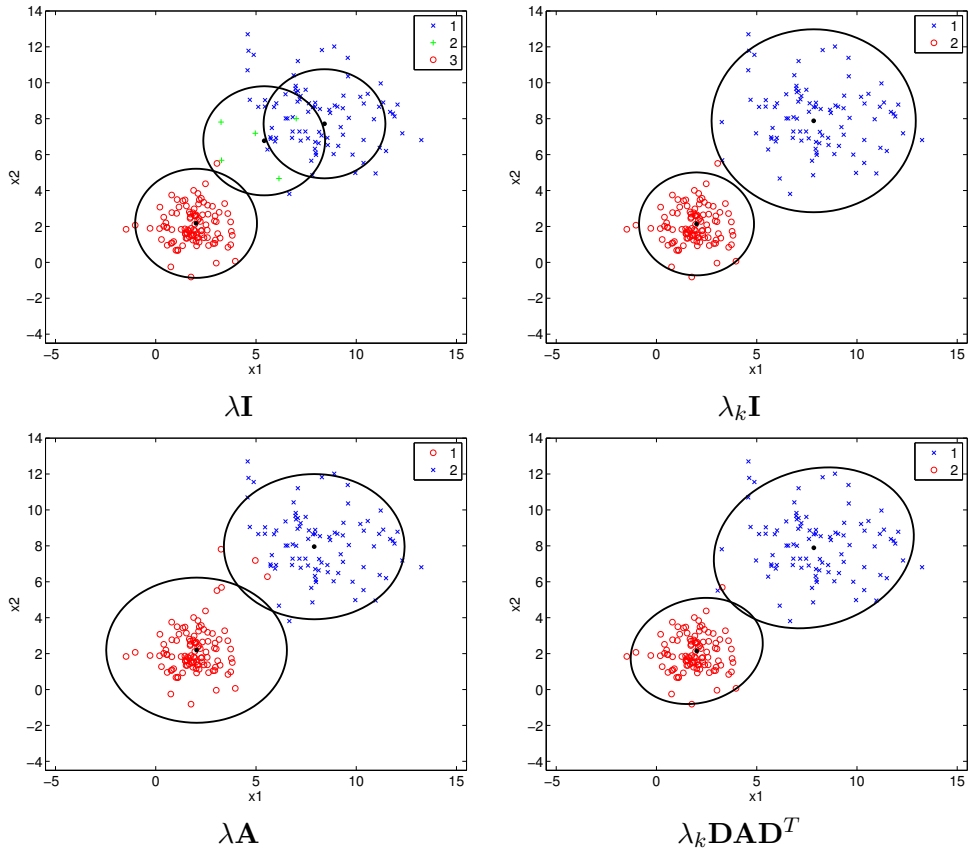


Figure 3.7: *The obtained partitions of the Gibbs sampling for the parsimonious GMMs over two component spherical dataset represented in Figure 3.3. The fourth hyperparameter setting of Table 5.12 is used.*

mation criteria. In this experiment, we compute the following information criteria: BIC, AIC, ICL and AWE. The corresponding results are shown in Figure 3.9. It shows that, for the Bayesian inference using Gibbs sampling, the values computed for the AWE criteria, descend also more sharply then the BIC, ICL or AIC criteria meaning a more decisive model selection for the parsimonious GMMs.

3.6 Conclusion

Up to here, the traditional Bayesian and non-Bayesian parametric mixture modeling approaches were discussed. In this chapter, we first described the general Bayesian GMM modeling, and then investigated the Bayesian parsimonious GMMs, which offer a great modeling flexibility. We focused on the inference using MCMC, and implemented, and assessed dedicated Gibbs

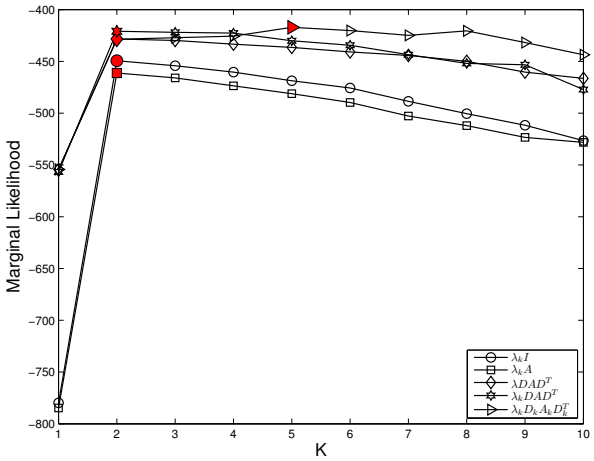


Figure 3.8: *Model selection using the Bayes Factors for the Old Faithful Geyser dataset. The parameters are estimated with Gibbs sampling.*

sampling algorithm. We provided a way to answer the main questions: how many components are needed and what is the best model structure to fit at best the data. The Bayes Factor, or some approximation of it have outlined to be one solution to this issue: the optimal number of components (e.g. clusters) and the best model structure (that is the eigenvalue decomposition of covariance matrix) for the parsimonious models.

However, this extra step, for selecting the number of clusters, can be omitted by using one alternative approach, that treats this problem of model selection in a different way (Hjort et al., 2010). This is the Bayesian non-parametric (BNP) alternative. In the next chapter, the Bayesian non-parametric (BNP) model that provides a flexible alternative model to the Bayesian, and non-Bayesian, parametric mixture models, is introduced. We propose new Bayesian non-parametric mixture models by introducing parsimony for the standard Bayesian non-parametric approach.

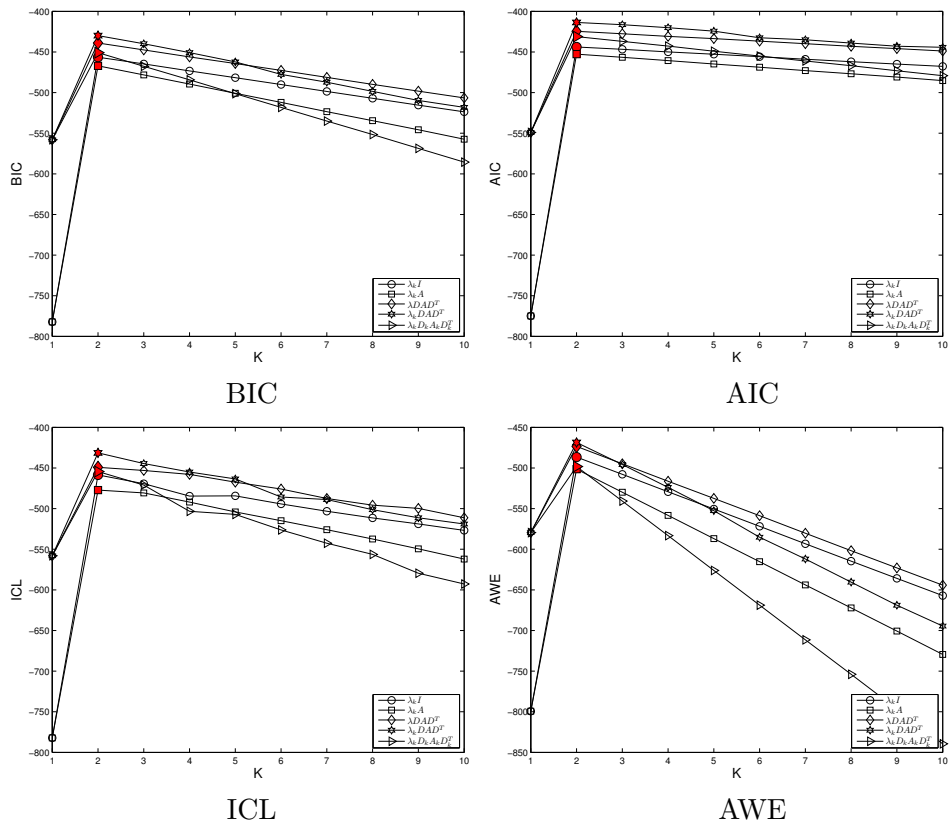


Figure 3.9: *Model selection for the Old Faithful Geyser dataset by using BIC (top left), AIC (top right), ICL (bottom left), AWE (bottom right). The models are estimated by Gibbs sampling.*

- Chapter 4 -

Dirichlet Process Parsimonious Mixtures (DPPM)

Contents

4.1 Introduction

In the previous chapters, we addressed the problem of model-based clustering by fitting finite Gaussian mixture, first in a MLE framework by relying on the EM algorithm, and then by mainly Bayesian MCMC sampling. We therefore tried to answer the question of how to fit at best a model to a complex data structure, while providing the well suited number of mixture components, and the more adapted model structure, in particular for the Bayesian parametric parsimonious GMMs. The analysis scheme was mainly two fold, that is, the selection of a model from previously estimated model candidates with different model structures, and in particular with different number of components. However, often, in a complex data, the scientist may not well select the good models (by supposing a bad number of components (clusters)) to fit the data, and as a result, they may not be well adapted.

In this chapter we will tackle the problem of model-based clustering, this is the one of Bayesian non-parametric mixture modeling. We discuss the Bayesian non-parametric approach of the Gaussian mixture model. We also propose a new Bayesian non-parametric (BNP) formulation of the parsimonious Gaussian mixture models, with the eigenvalue decomposition of the group covariance matrix for each component mixture which has proven its flexibility in cluster analysis in the parametric case (Banfield and Raftery, 1993; Bensmail and Meulman, 2003; Bensmail et al., 1997; Bensmail, 1995; Bensmail and Celeux, 1996; Celeux and Govaert, 1995; Fraley and Raftery, 2002, 2007a, 2005).

We develop new Dirichlet Process mixture models with parsimonious covariance structure, which results in Dirichlet Process Parsimonious Mixtures (DPPM). DPPMs represent a Bayesian non-parametric formulation of both the non-Bayesian and the Bayesian parsimonious Gaussian mixture models (Bensmail and Meulman, 2003; Bensmail et al., 1997; Bensmail, 1995; Bensmail and Celeux, 1996; Celeux and Govaert, 1995; Fraley and Raftery, 2002, 2007a, 2005). The proposed DPPM models are Bayesian parsimonious mixture models with a Dirichlet Process prior and thus provide a principled way to overcome the issues encountered in the parametric Bayesian and non-Bayesian case and allow to automatically and simultaneously infer the model parameters and the optimal model structure from the data, from different models, going from simplest spherical ones to the more complex standard general one. We develop a Gibbs sampling technique for maximum a posteriori (MAP) estimation of the various models and provide an unifying framework for model selection and models comparison by using namely Bayes factors, to simultaneously select the optimal number of mixture components and the best parsimonious mixture structure. The proposed DPPM are therefore more flexible in terms of modeling and their use in clustering, and automatically infer the number of clusters from the data.

We first provide an account on BNP mixture modeling in the next section and introduce some concepts needed for the developed Dirichlet Process parsimonious mixture models. Also, in order to validate our new approach, in the next chapter we discuss an experimental protocol for the generated data sets and real world data sets. The Bayesian parametric approach experimental protocol was also investigated in this chapter to make comparisons with the new proposed Dirichlet Process Parsimonious mixture approach.

4.2 Bayesian non-parametric mixtures

The Bayesian and non-Bayesian finite mixture models, described in the previous chapters, are in general parametric and may not be well adapted to represent complex and realistic data sets. Recently, the Bayesian non-parametric (BNP) formulation of mixture models, that goes back to Ferguson (1973) and Antoniak (1974), took much attention as a non-parametric alternative for formulating mixtures. The Bayesian non-parametric approach fits a mixture model to the data in a one fold scheme, rather then comparing multiple models that vary in complexity (regarding mainly the number of mixture components in a two fold strategy). The BNP methods (Hjort et al., 2010; Navarro et al., 2006; Orbanz and Teh, 2010; Robert, 1994; Teh and Jordan, 2010) have indeed recently become popular due to their flexible modeling capabilities and advances in inference techniques, in particular for mixture models, by using namely MCMC sampling techniques (Neal, 2000; Rasmussen, 2000) or variational inference ones (Blei and Jordan, 2006). BNP methods for clustering (Hjort et al., 2010; Robert, 1994), including Dirichlet Process Mixtures (DPM) and Chinese Restaurant Process (CRP) mixtures (Antoniak, 1974; Ferguson, 1973; Pitman, 1995; Samuel and Blei, 2012; Wood and Black, 2008) represented as Infinite Gaussian Mixture Models (IGMM) Rasmussen (2000), provide a principled way to overcome the issues encountered in standard model-based clustering and classical Bayesian mixtures for clustering. BNP mixtures for clustering are fully Bayesian approaches that offer a principled alternative to jointly infer the number of mixture components (i.e clusters) and the mixture parameters, from the data, rather than in a two-stage approach as in standard Bayesian and non-Bayesian model-based clustering (Hjort et al., 2010; Rasmussen, 2000; Samuel and Blei, 2012). By using general processes as priors, they allow to avoid the problem of singularities and degeneracies of the MLE, and to simultaneously infer the optimal number of clusters from the data, in a one-fold scheme, rather than in a two-fold approach as in standard model-based clustering. They also avoid assuming restricted functional forms and thus allow the complexity and accuracy of the inferred models to grow as more data is observed. They represent a good alternative to the difficult problem of model selection in parametric mixture models.

From the generative point of view, the Bayesian non-parametric mixture assumes that the observed data are governed by an infinite number of components, but only a finite number of them does actually generate the data. The term of non-parametric here does not mean that there are no parameters, but rather means that the number of parameters grows with the number of data, in such a way that only a (small) finite number of clusters will be actually active. 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. Dirichlet Process (Antoniak, 1974; Ferguson, 1973; Samuel and Blei, 2012) are commonly used as prior for the Bayesian non-parametric models.

In order to understand better the generative process for the Bayesian non-parametric mixture models, in the next section, we discuss the Dirichlet Process and some of it's possible equivalence as the Polya Urn scheme (Blackwell and MacQueen, 1973; Hosam, 209), the Stick Breaking construction (Sethuraman, 1994), and the Chinese Restaurant Process (CRP) (Aldous, 1985; Pitman, 2002; Samuel and Blei, 2012). Then the Dirichlet Process mixture models and the generative process are introduced.

4.2.1 Dirichlet Processes

Bayesian non-parametric priors were developed (Ferguson, 1974; Freedman, 1965), however in this work we are mostly focused on the Dirichlet Process prior.

Suppose a measure space Θ with a probability distribution on that space G_0. A Dirichlet Process (DP) (Ferguson, 1973) is a stochastic process, defining distribution over distributions, and has two parameters: the scalar concentration parameter $\alpha > 0$ and the base measure G_0. Each draw from a Dirichlet Process is a random probability measure G over Θ, such that for a finite measurable partition $(A_1, \dots A_k)$ of Θ, the random vector $(G(A_1), \dots G(A_k))$ is distributed as a finite dimensional Dirichlet distribution with parameters $(\alpha G_0(A_1), \dots, \alpha G_0(A_k))$, that is:

$$(G(A_1), \dots G(A_k)) \sim \mathrm{Dir}(\alpha G_0(A_1), \dots, \alpha G_0(A_k)).$$

We note that G is distributed according to a Dirichlet Process with base distribution G_0 and the concentration parameter α, that is:

$$G \sim \mathrm{DP}(\alpha, G_0). \tag{4.1}$$

The Dirichlet Process in Equation (4.1), has therefore two parameters: the base measure G_0, which can be interpreted as the mean of the DP, meaning that, the expected measure, for any set $A \subset \Theta$, of the random sample of the Dirichlet process and equals to $E[G(A)] = G_0(A)$, and the concentration

parameter α. This parameter can be interpreted as an inverse variance $V[G(A)] = G_0(A)(1 - G_0(A))/\alpha + 1$. Larger the α parameter is, smaller the variance will be, and the Dirichlet Process will concentrate more of it's mass on the mean. As a result, this parameter controls the number of clusters that appear in the data. The parameter α is also named the strength parameter or mass parameter (Teh, 2010).

The Dirichlet process has very interesting properties for the clustering perspective, as it provides the possibility of estimating the mixture components and respectively their number from the data. Assume there is a parameter $\tilde{\boldsymbol{\theta}}_i$ following a distribution G, that is $\tilde{\boldsymbol{\theta}}_i | G \sim G$. Modeling with DP means that we assume that the prior over G is a DP, that is, G is itself generated from a DP $G \sim \mathrm{DP}(\alpha, G_0)$. Thus, generating parameters and thus distributions from a DP can be summarized by the following generative process:

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_i | G &\sim G, \ \forall i \in 1, \ldots, n, \\ G | \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \cdot \end{aligned} \tag{4.2}$$

Note that the resulting random distribution G drawn from the Dirichlet Process, is from the same space as the base measure G_0. For example, if G_0 is univariate Gaussian then G will result a distribution over $\mathbb{R}$, as well as G is multivariate Gaussian if the base measure G_0 is a multivariate Gaussian distribution.

One of the main property of DP says that, draws from DP are discrete. With this consideration, there is a strictly positive probability that multiple observations $\tilde{\boldsymbol{\theta}}_i$, takes identical values within the set $(\tilde{\boldsymbol{\theta}}_1, \cdots, \tilde{\boldsymbol{\theta}}_n)$. The DP therefore places its probability mass on a countability infinite collection of points, also called atoms $\boldsymbol{\theta}_k, \forall\ k = 1, 2, \ldots$, that is an infinite mixture of Dirac deltas (Ferguson, 1973; Samuel and Blei, 2012; Sethuraman, 1994):

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k} \quad \boldsymbol{\theta}_k | G_0 \sim G_0, \ k = 1, 2, ..., \tag{4.3}$$

where π_k represents the probability assigned to the kth atom which satisfy $\sum_{k=1}^{\infty} \pi_k = 1$, and $\boldsymbol{\theta}_k$ is the location or value of that component (atom). These atoms are drawn independently from the base measure G_0. Hence, according to the DP process, the generated parameters $\tilde{\boldsymbol{\theta}}_i$ exhibit a clustering property, that is, they share repeated values with positive probability where the unique values of $\tilde{\boldsymbol{\theta}}_i$ shared among the variables are independent draws for the base distribution G_0 (Ferguson, 1973; Samuel and Blei, 2012). The Dirichlet process therefore provides a very interesting approach for clustering perspective, when we do not have a fixed number of clusters, in other words having an infinite mixture saying K tends to infinity.

Different representations of the Dirichlet Process can be found in the literature. We describe the main representations, that is, the Pólya Urn

representation, the Chinese Restaurant Process and the Stick-breaking construction. These representations can then be used for the developed Dirichlet Process mixtures models.

4.2.2 Pólya Urn representation

Suppose we have a random distribution G drawn from a DP followed by repeated draws $(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n)$ from that random distribution, Blackwell and MacQueen (1973) introduced a Pólya urn representation of the joint distribution of the random variables $(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n)$, that is

$$p(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n) = p(\tilde{\boldsymbol{\theta}}_1)p(\tilde{\boldsymbol{\theta}}_2|\tilde{\boldsymbol{\theta}}_1)p(\tilde{\boldsymbol{\theta}}_3|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) \ldots p(\tilde{\boldsymbol{\theta}}_n|\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_{n-1}), \quad (4.4)$$

which is obtained by marginalizing out the underlying random measure G:

$$p(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_n|\alpha, G_0) = \int \left(\prod_{i=1}^{n} p(\tilde{\boldsymbol{\theta}}_i|G) \right) \mathrm{d}p(G|\alpha, G_0) \quad (4.5)$$

and results in the following Pólya urn representation for the calculation of the predictive terms of the joint distribution (4.4):

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_i|\tilde{\boldsymbol{\theta}}_1, \ldots \tilde{\boldsymbol{\theta}}_{i-1} &\sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{j=1}^{i-1} \frac{1}{\alpha + i - 1} \delta_{\tilde{\boldsymbol{\theta}}_j} & (4.6) \\ &\sim \frac{\alpha}{\alpha + i - 1} G_0 + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1} \delta_{\boldsymbol{\theta}_k} & (4.7) \end{aligned}$$

where $K_{i-1} = \max\{z_j\}_{j=1}^{i-1}$ is the number of clusters after $i-1$ samples, n_k denotes the number of times each of the parameters $\{\boldsymbol{\theta}_k\}_{k=1}^{\infty}$ occurred in the set $\{\tilde{\boldsymbol{\theta}}\}_{i=1}^{n}$.

The DPPM model implements the Chinese Restaurant process representation of the Dirichlet Process, that provides a principled way to overcome the issues in standard model-based clustering and classical Bayesian mixtures for clustering.

4.2.3 Chinese Restaurant Process (CRP)

Consider the unknown cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$, where each value z_i is an indicator random variable that represents the label of the unique value $\boldsymbol{\theta}_{z_i}$ of $\boldsymbol{\theta}'_i$ such that $\boldsymbol{\theta}'_i = \boldsymbol{\theta}_{z_i}$ for all $i \in \{1, \ldots, n\}$. 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 $(z_1, \ldots, z_n)$:

$$p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1) \ldots p(z_n|z_1, z_2, \ldots, z_{n-1}) \cdot \quad (4.8)$$

From the Pólya urn distribution (Equation (4.7)), each predictive term of the joint distribution (Equation (4.8)) is given by the following:

$$p(z_i = k|z_1, ..., z_{i-1}; \alpha) = \frac{\alpha}{\alpha + i - 1}\delta(z_i, K_{i-1} + 1) + \sum_{k=1}^{K_{i-1}} \frac{n_k}{\alpha + i - 1}\delta(z_i, k)\cdot \tag{4.9}$$

where $n_k = \sum_{j=1}^{i-1} \delta(z_j, k)$ is the number of indicator random variables taking the value k, and $K_{i-1} + 1$ is the previously unseen value. 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. The distribution on partitions induced by the sequence of conditional distributions in Equation (4.9) is commonly referred to as the Chinese Restaurant Process (CRP).

The CRP name relates the following interpretation. 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 ($k \leq K_{i-1}$ being a previously occupied table), and may choose a new table ($k > K_{i-1}$, k being a new table to be occupied) with a probability proportional to a small positive real number α, which represents the CRP concentration parameter.

In clustering with the CRP, customers correspond to data points and tables correspond to clusters. A representation of the Chinese Restaurant Process can be seen in the Figure 4.1. In CRP mixture, the prior

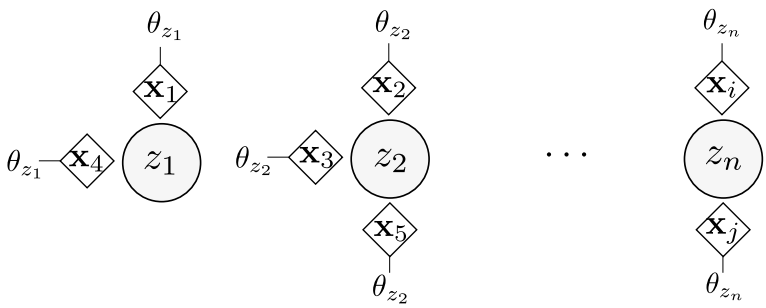


Figure 4.1: *A Chinese Restaurant Process representation.*

CRP$(z_1, \ldots, z_{i-1}; \alpha)$ is completed with a likelihood with parameters $\boldsymbol{\theta}_k$ for 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 therefore be summarized according

to the following generative process:

$$\begin{aligned} z_i &\sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \\ \boldsymbol{\theta}_{z_i}|G_0 &\sim G_0 \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim p(.|\boldsymbol{\theta}_{z_i}), \end{aligned} \tag{4.10}$$

where the CRP distribution is given by Eq. (4.8), G_0 is the base measure (that can be also seen as the prior distribution) and $p(\mathbf{x}_i|\boldsymbol{\theta}_{z_i})$ is a cluster-specific density. Two examples of draws from the CRP with 500 data points can be seen in Figure 4.2. One can see the difference when we vary the concentration parameter α. On left of Figure 4.2 $\alpha = 10$ and on right $\alpha = 1$. This clearly shows the property of the concentration parameter, that is, when it is higher, more tables (or components when modeling with the mixture model) will be generated. However, when α is small only a few number of tables (cluster) will be visited.

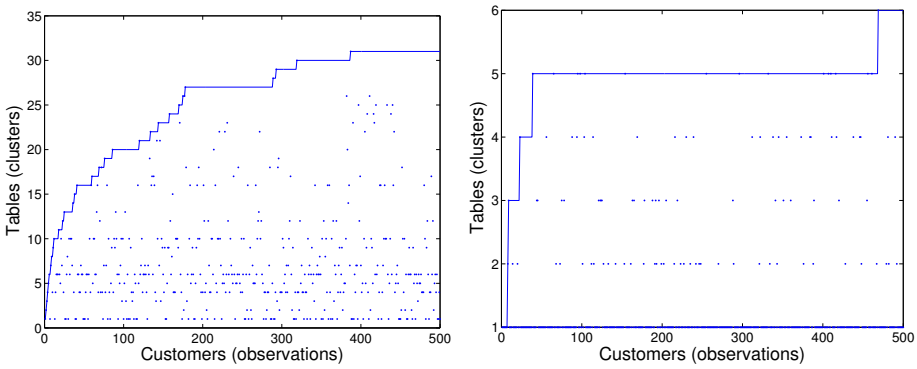


Figure 4.2: *A draw from a Chinese Restaurant Process sampling with* 500 *data points and* $\alpha = 10$ *(left) and* $\alpha = 1$ *(right). For* $\alpha = 10$, 31 *components are generated, and for* $\alpha = 1$ *only* 6 *components are visited.*

4.2.4 Stick-Breaking Construction

The fact that draws from the Dirichlet Process are discrete with probability 1 (Ferguson, 1973) is explicitly highlighted in the stick-breaking construction by (Sethuraman, 1994). The Stick-Breaking constructing is derived as follows. Suppose the base measure G_0 on the space Θ, it was showed that the random measure G can be defined as an infinite sum of weight point masses:

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\theta}_k},$$

where the Dirac $\delta_{\boldsymbol{\theta}_k}$ being the probability measure concentrated at $\boldsymbol{\theta}_k$, and π_k $\forall k = 1, 2, \ldots$ being the weights. In the Stick-Breaking construction

the weights are assumed to be sampled from the infinite sequence of beta distributions.

$$\pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1} (1 - \tilde{\pi}_l). \tag{4.11}$$

The independent sequence of the i.i.d random variables $(\tilde{\pi}_k)_{k=1}^{\infty}$ and $(\boldsymbol{\theta}_k)_{k=1}^{\infty}$ being sampled as:

$$\begin{aligned} \tilde{\pi}_k | \alpha, G_0 &\sim \text{Beta}(1, \alpha), \\ \boldsymbol{\theta}_k | \alpha, G_0 &\sim G_0, \end{aligned} \tag{4.12}$$

where the sequence $(\pi_k)_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$ with probability 1. The stick breaking process is noted by $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$ ("GEM" stands for Griffiths, Engen, and McCloskey (Pitman, 2002; Teh, 2010)). Example of samples for the stick breaking process is showed in Figure 4.3 with respectively $\alpha = 1, 2$ and 5.

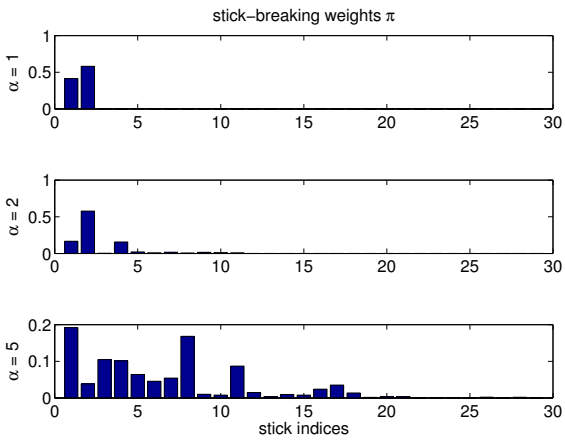


Figure 4.3: *A Stick-Breaking Construction sampling with $\alpha = 1$ (top), $\alpha = 2$ (middle) and $\alpha = 5$ (bottom).*

Because of it's richness, computation ease and interpretability, the Dirichlet Process (DP) is one of the most important random probability measures that are mostly used for the Bayesian non-parametric models. The resulting Bayesian non-parametric mixture using DP prior is called the Dirichlet Process mixture models. In the next section, we rely on the DP formulation of mixture models to develop DP parsimonious mixture models.

4.2.5 Dirichlet Process Mixture Models

The idea of DP mixture models is to incorporate the Dirichlet Process prior into the Bayesian mixture model shown in Equation (3.1). Clustering with

DP, adds a third step to the DP generative model (4.2), that is, the random variables $\mathbf{x}_i$, given the distribution parameters $\tilde{\boldsymbol{\theta}}_i$ which are generated from a DP, are generated from a conditional distribution $p(.|\tilde{\boldsymbol{\theta}}_i)$.

This is the DP Mixture model (DPM) (Antoniak, 1974; Escobar, 1994; Samuel and Blei, 2012; Wood and Black, 2008). The generative process DPM, is therefore given by:

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \tilde{\boldsymbol{\theta}}_i|G &\sim G \\ \mathbf{x}_i|\tilde{\boldsymbol{\theta}}_i &\sim p(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}_i) \end{aligned} \tag{4.13}$$

where $p(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}_i)$ is a cluster-specific density. Figure 4.4 shows the graphical representation of the DPM model.

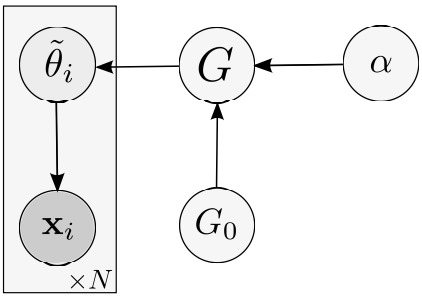


Figure 4.4: *Probabilistic graphical model representation of the Dirichlet Process Mixture Model (DPM). The data are supposed to be generated from the distribution $p(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}_i)$ parametrized with $\tilde{\boldsymbol{\theta}}_i$ which are generated from a DP.*

When K tends to infinity, it can be shown that the finite Bayesian mixture model (4.15) converges to a Dirichlet process mixture model (Ishwaren and Zarepour, 2002; Neal, 2000; Rasmussen, 2000). The Dirichlet process has a number of properties which make inference based on this non-parametric prior computationally tractable. It has a interpretation in term of the CRP mixture (Pitman, 2002; Samuel and Blei, 2012). It has the property that random parameters drawn from a DP exhibit a clustering property, which connects the DP to the CRP. Consider a random distribution drawn from DP $G \sim DP(\alpha, G_0)$, followed by a repeated draws from that random distribution $\tilde{\boldsymbol{\theta}}_i \sim G$, $\forall i \in 1, \ldots, n$. The structure of shared values defines a partition of the integers from 1 to n, and the distribution of this partition is a CRP (Ferguson, 1973; Samuel and Blei, 2012). The Chinese Restaurant process construction used in the Infinite Gaussian mixture model introduced by Rasmussen (2000), where the cluster specific density $p(\mathbf{x}_i|\tilde{\boldsymbol{\theta}}_i)$ was considered to be a univariate normal density.

4.2.6 Infinite Gaussian Mixture Model and the CRP

Rasmussen (2000) developed the infinite mixture of the univariate GMMs, defining Normal-Gamma prior distribution as base measure (prior) over the corresponding mixture components, that is the mean μ_k and the variance σ_k^2 for component k. However, this work focuses on the multivariate data, as in Wood and Black (2008); Wood et al. (2006). Thus, the base measure G_0 may be a multivariate normal Inverse-Wishart conjugate prior distribution as in Wood and Black (2008); Wood et al. (2006).

$$G_0 = \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0)\mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0), \tag{4.14}$$

where $(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$ are the Bayesian Gaussian mixture hyperparameters discussed in Section 3.3.

The generative process for the infinite Gaussian mixture model based on the Chinese Restaurant Process (CRP) can be summarized as:

$$\begin{aligned} z_i|\alpha &\sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha), \\ \boldsymbol{\mu}_{z_i}|\boldsymbol{\mu}_0, \kappa_0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \kappa_0), \\ \boldsymbol{\Sigma}_{z_i}|\boldsymbol{\Lambda}_0, \nu_0 &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0), \\ \mathbf{x}_i|\boldsymbol{\theta}_{z_i} &\sim \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}). \end{aligned} \tag{4.15}$$

Figure 4.5 shows the probabilistic graphical model for the Chinese Restaurant Process mixture model. Note that, in the Dirichlet Process mixture

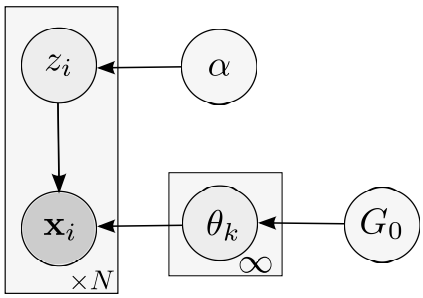


Figure 4.5: *Probabilistic graphical model for Dirichlet Process mixture model using the Chinese Restaurant Process construction.*

representation using CRP, the independence of the labels and the mixture parameters are made explicitly apart. The data partition results from the CRP, while the model parameters are drawn from the base measure, that is, the Normal inverse-Wishart distribution followed by generating the data from the cluster specific density, for example a multivariate Gaussian distribution in the GMM case.

4.2.7 Learning the Dirichlet Process models

Given n observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ modeled by the Dirichlet process mixture model (DPM), the aim is to infer the parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$,

the number K of latent clusters underlying the observed data and the latent cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$.

The Dirichlet Process mixture models can not be analytically estimated. This is performed by sampling inference techniques like MCMC sampling methods, that are easily adapted to the non-parametric models. Here we investigate the Gibbs sampling approach of the MCMC. This can be performed similarly as in the Bayesian parametric mixture models described in the previous chapter. The main idea of this sampling approach, is to upgrade the model parameters, including the cluster labels, conditioned on the rest of the model parameters and the observed data. Conjugate priors are used in this work, however, we mention that in literature one can found developed MCMC algorithms with non-conjugate priors on the DPM models Green and Richardson (2001); Görür and Edward Rasmussen (2010); Maceachern (1994).

Given an initial mixture parameters $\boldsymbol{\theta}^{(0)}$, and a prior over the missing labels $\mathbf{z}$ (here a conjugate Chinese Restuarant Process prior), the Gibbs sampler, instead of estimating the missing labels $\mathbf{z}^{(t)}$, simulates them from their posterior distribution $p(\mathbf{z}^{(t)}|\mathbf{X}, \boldsymbol{\theta}^{(t)})$ at each iteration t. Recall that the posterior is obtained by combining the prior with the likelihood. So, the cluster labels z_i are sampled from the posterior distributions given by:

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \boldsymbol{\Theta}, \alpha) \propto p(\mathbf{x}_i|z_i; \boldsymbol{\Theta})p(z_i|\mathbf{z}_{-i}; \alpha) \tag{4.16}$$

where $\mathbf{z}_{-i} = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)$, and $p(z_i|\mathbf{z}_{-i}; \alpha)$ is the prior predictive distribution which corresponds to the CRP distribution computed as in Equation (4.9). Then, given the completed data and the prior distribution $p(\boldsymbol{\theta})$ over the mixture parameters, the Gibbs sampler generates the mixture parameters $\boldsymbol{\theta}^{(t+1)}$ from the posterior distribution

$$p(\boldsymbol{\theta}_k|\mathbf{z}, \mathbf{X}, \boldsymbol{\Theta}_{-k}, \alpha; \mathcal{H}) \propto \prod_{i|z_i=k} p(\mathbf{x}_i|z_i = k; \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k; \mathcal{H}) \tag{4.17}$$

where $\boldsymbol{\Theta}_{-k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_{K_{i-1}})$ and $p(\boldsymbol{\theta}_k; \mathcal{H})$ is the prior distribution for $\boldsymbol{\theta}_k$, that is G_0, with $\mathcal{H}$ being the hyperparameters of the model. Generally, these hyperparameters are specified a priori by the user, and are not learned from the data. However, when using hierarchical methods they are sampled from the data, making the model more flexible and adaptive. This Bayesian sampling procedure produces an ergodic Markov chain of samples $(\boldsymbol{\theta}^{(t)})$ with a stationary distribution $p(\boldsymbol{\theta}|\boldsymbol{X})$. Therefore, after initial M burn-in steps in N Gibbs samples, the variables $(\boldsymbol{\theta}^{(M+1)}, ..., \boldsymbol{\theta}^{(N)})$, can be considered to be approximately distributed according to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$.

The DPM Gibbs sampling is derived in Pseudo-code 7.

Pseudo-code 7 can further be also simplified by integrating over the model parameters $\boldsymbol{\theta}$ and eliminating them from the Markov chain state,

Algorithm 6 Gibbs sampling for the conjugate priors DPM models

Inputs: Data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and # Gibbs samples

1: $t \leftarrow 1$
2: Initialize the Markov chain state that consists of the labels $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots z_n^{(t)})$ and the model parameters $\boldsymbol{\theta}_{\mathbf{z}^{(t)}}^{(t)}$.
3: **for** $t = 2, \ldots, \#$samples **do**
4: **for** $i = 1, \ldots, n$ **do**
5: Sample a cluster label $z_i^{(t)}$ from according to its posterior that is the product of the likelihood and the prior over the cluster label, that is a Chinese Restaurant Process prior distribution (see Equation (4.16)).
6: For $z_i^{(t)}$, sample the a new model parameter $\boldsymbol{\theta}_{\mathbf{z}^{(t)}}^{(t)}$ for this component according to the base distribution G_0 (see Equation (4.14)).
7: **end for**
8: Select the represented components K_{i-1} that is the number of unique values of $\boldsymbol{\theta}_{\mathbf{z}}^{(t)}$, thus removing the non representative model parameters from the modeling representation.
9: **for** $k = 1, \ldots, K_{i-1}$ **do**
10: Sample the parameters $\boldsymbol{\theta}_k^{(t)}$ from the posterior distribution conditional on the data, cluster labels and hyperparameters (see Equation (4.17)).
11: **end for**
12: **end for**

Outputs: The parameters vector chain of the mixture $\hat{\boldsymbol{\Theta}} = \{\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}, \ \forall t = 1, \ldots, n_s$.

thus the sampling procedure reduces only to sampling the indicator labels $\mathbf{z}$. This algorithm is known as Rao-Blackwellized MCMC sampling or collapsed Gibbs sampling (Andrieu et al., 2003; Casella and Robert, 1996; Görür, 2007; Neal, 2000; Sudderth, 2006; Wood, 2007). However, the need of estimating the model parameters in our developed parsimonious models, described in the next section, makes this case not appropriate of this work. We have therefore concentrated on the purpose of estimating all the mixture parameters as well as the hidden cluster indicators. The parsimonious models are discussed in the following section.

4.3 Chinese Restaurant Process parsimonious mixture models

We previously saw how finite parsimonious mixture models were derived from the finite mixture models framework. Clustering with parsimonious models gives different opportunities, like reducing the number of parameters to estimate in the model and giving different flexible models that control the clusters structure in the data. Thus, to take benefit of these advantages in the BNP framework, we develop parsimonious BNP models. We introduce infinite multivariate Gaussian mixture model with the Chinese Restaurant Process prior over the hidden labels $\mathbf{z}$. The parsimony considered in the eigenvalue decomposition of the covariance matrix is introduced for each model component. We name this approach the Dirichlet Process Parsimonious mixture (DPPM) model, that is equivalent to the Chinese Restaurant Process Parsimonious Mixture Models or more generally the Infinite Parsimonious Gaussian Mixture Models.

Suppose the Chinese Restaurant Process Mixture, where the metaphor of CRP is used to sample the labels. As in the Chinese Restaurant Process, the clients visiting the restaurant are social, so that the ith customer will sit 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 is given by:

$$\begin{aligned} p(z_i = k | z_1, ..., z_{i-1}) &= \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \\ &= \begin{cases} \frac{n_k}{i-1+\alpha} & \text{if} \quad k \leq K_{i-1} \\ \frac{\alpha}{i-1+\alpha} & \text{if} \quad k > K_{i-1} \end{cases} \end{aligned} \tag{4.18}$$

where $k \leq K_{i-1}$ is a previously occupied table and $k > K_{i-1}$, k is a new occupied table.

Suppose that, the data are Gaussian, then, the model parameters are sampled according to the base distribution G_0 that is a Normal distribution for the mean vector and an inverse-Wishart distribution for the covariance matrix.

We use the eigenvalue value decomposition described in section 2.4.3 which till now has been considered only in the case of parametric finite mixture model-based clustering (Banfield and Raftery, 1993; Celeux and Govaert, 1995), and Bayesian parametric finite mixture model-based clustering (Bensmail and Meulman, 2003; Bensmail et al., 1997; Fraley and Raftery, 2007a, 2005). Recall that for the GMM we have the following prior form:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\pi}|\boldsymbol{\alpha}) p(\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \kappa_0) p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \nu, \boldsymbol{\Lambda}_0)$$

where $(\boldsymbol{\alpha}, \boldsymbol{\mu}_0, \kappa_0, \nu, \boldsymbol{\Lambda}_0)$ are hyperparameters that can be tuned from the data. A common choice is to assume conjugate priors, that is Dirichlet distribution for the mixing proportions $\boldsymbol{\pi}$ as in Richardson and Green (1997) Ormoneit and Tresp (1998), and a multivariate normal Inverse-Wishart prior distribution for the Gaussian parameters, that is a multivariate normal for the means $\boldsymbol{\mu}$ and an Inverse-Wishart for the covariance matrices $\boldsymbol{\Sigma}$ as in Fraley and Raftery (2007a, 2005) Bensmail et al. (1997).

The used priors on the model parameters depend on the type of the parsimonious model (see Table 4.1). Thus, sampling the model parameters varies according to the considered parsimonious mixture model. Indeed, yet we investigated nine 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 4.1 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 Bensmail et al. (1997). The base distribution $G_0(\boldsymbol{\mu}_k)$ will be a normal distribution ($\mathcal{N}$) for all the models.

#	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	$\lambda\mathbf{A}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
4	$\lambda_k\mathbf{A}$	Diagonal	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}$
5	$\lambda\mathbf{DAD}^T$	General	$\mathcal{IW}$	$\boldsymbol{\Sigma} = \lambda\mathbf{DAD}^T$
6	$\lambda_k\mathbf{DAD}^T$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ_k and $\boldsymbol{\Sigma} = \mathbf{DAD}^T$
7	$\lambda\mathbf{DA}_k\mathbf{D}^{T*}$	General	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}_k$
8	$\lambda_k\mathbf{DA}_k\mathbf{D}^{T*}$	General	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}_k$
9	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	General	$\mathcal{IG}$	each diagonal element of $\lambda\mathbf{A}$
10	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	General	$\mathcal{IG}$	each diagonal element of $\lambda_k\mathbf{A}$
11	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^{T*}$	General	$\mathcal{IG}$ and $\mathcal{IW}$	λ and $\boldsymbol{\Sigma}_k = \mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$
12	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	General	$\mathcal{IW}$	$\boldsymbol{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

Table 4.1: *Considered Parsimonious GMMs via eigenvalue decomposition, the associated prior for the covariance structure and the corresponding number of free parameters where $\mathcal{I}$ denotes an inverse distribution, $\mathcal{G}$ a Gamma distribution and $\mathcal{W}$ a Wishart distribution.*

4.4 Learning the Dirichlet Process parsimonious mixtures using Gibbs sampling

Given n observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ modeled by the proposed Dirichlet process parsimonious mixture (DPPM), the aim is to infer the number K of latent clusters underlying the observed data, their parameters $\boldsymbol{\Psi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ and the latent cluster labels $\mathbf{z} = (z_1, \ldots, z_n)$. Note that, in DPPM, the components are Gaussian so $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ where the covariance takes the eigenvector parametrization, so that according to each parsimonious model we can have the following parameters: $\{\lambda_k, \mathbf{D}_k, \mathbf{A}_k\}$, representing respectively the volume, orientation and the shape for each cluster. These parameters can also be constrained, to be equal, for each of the component, obtaining that way a more parsimonious model.

In this section, we developed an MCMC Gibbs sampling technique, as in Neal (2000); Rasmussen (2000); Wood and Black (2008), to learn the proposed Bayesian non-parametric parsimonious mixture models. The first form of Gibbs sampler goes back to Geman and Geman (1984) and was proposed in a framework of Bayesian image restoration. A version very close to it was introduced by Tanner and Wong (1987) under the name of data augmentation for missing data problems, and was shown in Gelfand and Smith (1990) and Diebolt and Robert (1994). The idea of the Markov chain based on the Gibbs sampling relies on updating the parameters, the hyperparameters, and the cluster labels for the proposed model. Updating all these model variables are made according to their posterior distribution conditional on all other variables. A summary of such a method can be given as follows.

- Update the cluster labels conditional on the other indicators, all the parameters and hyperparameters of the model and the observed data.
- Update the mixture parameters: the mean vector and the covariance matrix taking the eigenvector decomposition, conditional on the observed data, class labels and the hyperparameters.
- Update the model hyperparameters, particularly the concentration hyperparameter α of the Dirichlet Process.

Sampling the hidden cluster labels The cluster labels z_i are sampled from the posterior distribution, which is given by:

$$p(z_i = k|\mathbf{z}_{-i}, \mathbf{X}, \boldsymbol{\Theta}, \alpha) \propto p(\mathbf{x}_i|z_i; \boldsymbol{\Theta})p(z_i|\mathbf{z}_{-i}; \alpha)$$

is calculated by multiplying the likelihood term $p(\mathbf{x}_i|z_i; \boldsymbol{\Theta})$ with the prior predictive distribution corresponding to the CRP distribution computed as in Equation (4.18). Here the likelihood term would be a Gaussian distribution $\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ where the specific family model: the spherical, the

diagonal or the general one, parametrizes the covariance matrices according to the eigenvector decomposition. Note that the likelihood term is given for each of the data point $\mathbf{x}_i$ that is associated to it's class label z_i, and according to the Dirichlet Process clustering property (Antoniak, 1974), when grouping equal parameters $\boldsymbol{\theta}_i$ we obtain the unique values that are the active components $\boldsymbol{\theta}_k$. That is, when we choose to assign a data point $\mathbf{x}_i$ to the existing components, or a new active component will be created by sampling according to the base distribution G_0 that will be conditioned on the eigenvalue decomposition of the covariance matrix.

Sampling the mixture parameters When the number of active components in the mixture is known, the Gibbs sampler consists therefore in sampling the mixture parameters from their posterior distribution. The posterior distribution for $\boldsymbol{\theta}_k$ given all the other variables is given by the product of the likelihood distribution and $p(\boldsymbol{\theta}_k; \mathcal{H})$ the prior distribution for $\boldsymbol{\theta}_k$, that is a conjugate base distribution G_0, with $\mathcal{H}$ the model hyperparameters.

$$p(\boldsymbol{\theta}_k|\mathbf{z}, \mathbf{X}, \boldsymbol{\Theta}_{-k}, \alpha; \mathcal{H}) \propto \prod_{i|z_i=k} p(\mathbf{x}_i|z_i = k; \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k; \mathcal{H})$$

where $\boldsymbol{\Theta}_{-k} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_{k-1}, \ldots, \boldsymbol{\theta}_{K_{i-1}})$ are all the active model parameters except the one that is sampled $\boldsymbol{\theta}_k$.

Sampling the concentration hyperparameter The number of mixture components in the models depends on the hyperparameter α of the Dirichlet Process (Antoniak, 1974). Therefore it is natural to sample this hyperparameter, to make the model more flexible, avoiding fixing it an arbitrary value for it. The method introduced by Escobar and West (1994) consists in sampling α hyperparameter, by assuming a prior Gamma distribution $\alpha \sim \mathcal{G}(a, b)$ with a shape hyperparameter $a > 0$ and scale hyperparameter $b > 0$. Then, a variable η is introduced and sampled conditionally on α and the number of clusters K_{i-1}, according to a Beta distribution $\eta|\alpha, K_{i-1} \sim \mathcal{B}(\alpha + 1, n)$. The resulting posterior distribution for the hyperparameter α is given by:

$$p(\alpha|\eta, K) \sim \vartheta_\eta \mathcal{G}\left(a + K_{i-1}, b - \log(\eta)\right) + (1 - \vartheta_\eta)\, \mathcal{G}\left(a + K_{i-1} - 1, b - \log(\eta)\right) \tag{4.19}$$

where the weights $\vartheta_\eta = \frac{a+K_{i-1}-1}{a+K_{i-1}-1+n(b-\log(\eta))}$. The retained solution is the one corresponding to the posterior mode of the number of mixture components, that is the one that appears the most frequently during the sampling.

The MCMC Gibbs sampling technique, to learn the proposed Bayesian non-parametric mixture models is derived in Pseudo-code 7.

Note that, the parameter vector is obtained by averaging the Gibbs samples for the partition that appears the most frequently during the sampling, after removing the burn-in period.

Algorithm 7 Gibbs sampling for the proposed DPPM

Inputs: Data set $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and # Gibbs samples

Initialize the model hyperparameters H.
Start with one cluster $K_1 = 1, \boldsymbol{\theta}_1 = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}$
for $t = 2, \ldots,$ #samples **do**
  for $i = 1, \ldots, n$ **do**
    for $k = 1, \ldots, K_{i-1}$ **do**
      if $(n_k = \sum_{i=1}^{N} z_{ik}) - 1 = 0$ **then**
        Decrease $K_{i-1} = K_{i-1} - 1$; let $\{\boldsymbol{\theta}^{(t)}\} \leftarrow \{\boldsymbol{\theta}^{(t)}\} \setminus \boldsymbol{\theta}_{z_i}$
      end if
    end for
    Sample a cluster label $z_i^{(t)}$ from the posterior:
    $p(z_i|\mathbf{z}_{\setminus z_i}, \mathbf{X}, \boldsymbol{\theta}^{(t)}, H) \propto p(\mathbf{x}_i|z_i, \boldsymbol{\theta}^{(t)})\mathrm{CRP}(\mathbf{z}_{\setminus z_i}; \alpha)$
    if $z_i^{(t)} = K_{i-1} + 1$ **then**
      Increase $K_{i-1} = K_{i-1} + 1$ (We get a new cluster) and sample a new cluster parameter $\boldsymbol{\theta}_{z_i}^{(t)}$ from the conjugate prior distribution $\mathcal{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$.
    end if
  end for
  for $k = 1, \ldots, K_{i-1}$ **do**
    Sample the parameters $\boldsymbol{\theta}_k^{(t)}$ from the posterior distribution.
  end for
  Sample the hyperparameter $\alpha^{(t)} \sim p(\alpha^{(t)}|K_{i-1})$ from the posterior (4.19)
  $\mathbf{z}^{(t+1)} \leftarrow \mathbf{z}^{(t)}$
end for

Outputs: The parameters vector chain of the mixture $\hat{\boldsymbol{\Theta}} = \{\boldsymbol{\pi}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}, \ \forall t = 1, \ldots, n_s$.

Complexity of the algorithm The method complexity is mainly related to the label z_i and model parameters $\boldsymbol{\theta}_i$ simulations, therefore it depends on the number of components or classes in data and the dimension of model parameters. Therefore, the complexity of each Gibbs sampler is proportional to the actual number of components (active components K_{i-1} being estimated automatically, as the data is learned), and randomly varies from one iteration to another, depending on the posterior distribution of the number of classes. Asymptotically, K tends to $\alpha \log(n)$ when n tends to infinity (Antoniak, 1974). Therefore, each sampler requires $O(\alpha n \log(n))$ operations for sampling the class labels z_i. The parameter simulation (the mean vector and the covariance matrix), requires in turn, in the worst case (when the covariance matrix takes the full mode) approximatively $O\left(\alpha \log(n)\ d^3\right)$ that

gives us a total complexity equal to $O\left(\alpha \log(n)\left(n + d^3\right)\right)$.

Label switching problem Compared to the frequentist case, in particular due to the label switching problem when simulating the label indicators does not effect the likelihood and the goodness of the model remains Redner and Walker (1984), the problem of label switching has to be addressed during the Bayesian inference, particularly in the MCMC techniques, when the prior distribution is symmetric in the components of the mixture. This phenomenon can produce unexpected results when label switching appears during the MCMC samples. To deal with this problem, different strategies were discussed in the literature.

One of the simplest way to deal with label switching is to use a constraint on the model parameters, so that the MCMC algorithm will be forced to use a unique labeling. For example suppose the model parameters $\theta_1, \ldots, \theta_K$. One possible constraint is to enforce an increasing order on the parameters like $\theta_1 < \ldots < \theta_K$. This strategy is used in Marin et al. (2005); Richardson and Green (1997). However, Celeux et al. (1999) showed that using constrains on model parameters to deal with label switching can lead to unsatisfactory result.

Celeux (1998) recommended to deal with the label switching problem without using any constrains on the parameters and then using a clustering-like algorithm at the end of the MCMC sampling when component label switchings appear. A similar approach was used by Stephens (1999).

So what is suggested is that either to relabel the samples upon a visual inspection or, what is suggested here, to cluster the obtained Gibbs samples and to see when the label switching appears in order to possibly relabel the samplers as suggested by Celeux (1998); Stephens (1999).

Model Selection and comparison for the DPPM This section provides the used strategy for model selection and comparison, that is the selection of the best model from different parsimonious models of DPPM. We use Bayes factors, described in Section 3.5.9. We approximate the marginal likelihood by Laplace-Metropolis approximation that gives appropriate results for the parsimonious models that we assume in this work. We note that, in the proposed DPPM models, as the number of components K is itself a parameter in the model and is changing during the sampling, which leads to parameters with different dimension, we compute the Hessian matrix $\hat{\mathbf{H}}$ in Equation (3.24) by taking the posterior samples corresponding to the posterior mode of K. We performed experiments over the simulated and real datasets in order to validate our Dirichlet Process Parsimonious Mixture approach. The detailed results for the model selection with the Bayes Factor is discussed in the next chapter.

4.5 Conclusion

In this chapter we presented Bayesian non-parametric parsimonious mixture models for clustering. It is based on an infinite Gaussian mixture with an eigenvalue decomposition of the cluster covariance matrix and a Dirichlet Process, or by equivalence a Chinese Restaurant Process prior. This 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 also proposed a Bayesian model selection an comparison framework to automatically select, the best model, with the best number of components, by using Bayes factors.

In the next chapter we investigate experiments over the simulated and real world data sets.

- Chapter 5 -

Application on simulated data sets and real-world data sets

Contents

5.1 Introduction

This chapter is dedicated to an experimental study of the proposed models. We performed experiments on both simulated and real data in order to evaluate our proposed DPPM models. We assess their flexibility in terms of modeling, their use for clustering and inferring the number of clusters from the data. We show how the proposed DPPM approach is able to automatically and simultaneously select the best model with the optimal number of clusters by using the Bayes factors, which is used to evaluate the results. We also perform comparisons with the finite model-based clustering approach (as in Bensmail et al. (1997); Fraley and Raftery (2007a)), which will be abbreviated as PGMM approach. We also use the Rand index to evaluate and compare the provided partitions, and the misclassification error rate when the number of estimated components equals the actual one.

For the simulations, we consider several situations of simulated data, from different models, and with different levels of cluster separations, in order to assess the efficiency of the proposed approach to retrieved the actual partition with the actual number of clusters. We also assess the stability of our proposed DPPMs models regarding the choice of the hyperparameters values, by considering several situations and varying them. Then, we perform experiments on several real data sets and provide numerical results in terms of comparisons of the Bayes factors (via the log marginal likelihood values) and as well the Rand index and the misclassification error rate for data sets with known actual partition. In the experiments, for each of the compared approaches and for each model, each Gibbs is run ten times with different initializations. Each Gibbs run generates 2000 samples for which 100 burn-in samples are removed. The solution corresponding to the highest Bayes factor, of those ten runs, is then selected.

5.2 Simulation study

5.2.1 Varying the clusters shapes, orientations, volumes and separation

In this experiment, we apply the proposed models on simulated data simulated according to different models, and with different level of mixture separation, going from poorly separated mixtures to very-well separated mixtures. To simulate the data, we first consider an experimental protocol close to the one used by Celeux and Govaert (1995) where the authors considered the parsimonious mixture estimation within a MLE framework. This therefore allows to see how do the proposed Bayesian non-parametric DPPM perform compared to the standard parametric non-Bayesian one. We note however that in Celeux and Govaert (1995) the number of com-

ponents was known a priori and the problem of estimating the number of classes was not considered. We have performed extensive experiments involving all the models and many Monte Carlo simulations for several data structure situations. Given the variety of models, data structures, level of separation, etc, it is not possible to display all the results in the paper. We choose to perform in the same way as in the standard paper Celeux and Govaert (1995) by selecting the results display, for the experiments on simulated data, fo six models of different structures. The data are generated from a two component Gaussian mixture in $\mathbb{R}^2$ with 200 observations. The six different structures of the mixture that have been considered to generate the data are: two spherical models: $\lambda\mathbf{I}$ and $\lambda_k\mathbf{I}$, two diagonal models: $\lambda\mathbf{A}$ and $\lambda_k\mathbf{A}$ and two general models $\lambda\mathbf{DAD}^T$ and $\lambda_k\mathbf{DAD}^T$. Table (5.1) shows the considered model structures and the respective model parameter values used to generate the data sets. Let us recall that the variation in

Model	Parameters values
$\lambda\mathbf{I}$	$\lambda = 1$
$\lambda_k\mathbf{I}$	$\lambda_k = \{1, 5\}$
$\lambda\mathbf{A}$	$\lambda = 1;\ \mathbf{A} = \mathrm{diag}(3, 1/3)$
$\lambda_k\mathbf{A}$	$\lambda_k = \{1, 5\};\ \mathbf{A} = \mathrm{diag}(3, 1/3)$
$\lambda\mathbf{DAD}^T$	$\lambda = 1;\ \mathbf{D} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}; \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$
$\lambda_k\mathbf{DAD}^T$	$\lambda_k = \{1, 5\};\ \mathbf{D} = \begin{bmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2}; \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$

Table 5.1: *Considered two-component Gaussian mixture with different structures.*

the volume is related λ, the variation of the shape is related to $\mathbf{A}$ and the variation of the orientation is related to $\mathbf{D}$. Furthermore, for each type of model structure, we consider three different levels of mixture separation, that is: poorly separated, well separated, and very-well separated mixture. This is achieved by varying the following distance between the two mixture components $\varrho^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. We consider the values $\varrho = \{1, 3, 4.5\}$. As a result, we obtain 18 different data structures with poorly ($\varrho = 1$), well ($\varrho = 3$) and very well ($\varrho = 4.5$) separated mixture components. As it is difficult to show the figures for all the situations and those of the corresponding results, in Figure 5.1, we show for three models with equal volume across the mixture components, different data sets with varying level of mixture separation. Respectively, in Figure 5.2, we show for the models with varying volume across the mixture components, different data sets with varying level of mixture separation.

We compare the proposed DPPM to the parametric PGMM approach in model-based clustering (Bensmail et al., 1997; Bensmail, 1995; Bensmail and Celeux, 1996), for which the number of mixture components was varying in

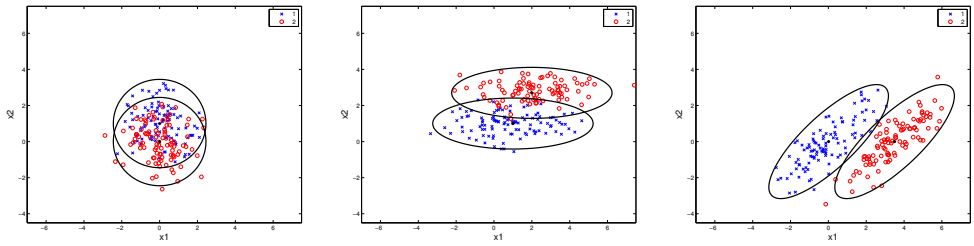


Figure 5.1: *Examples of simulated data with the same volume across the mixture components: spherical model $\lambda\mathbf{I}$ with poor separation (left), diagonal model $\lambda\mathbf{A}$ with good separation (middle), and general model $\lambda\mathbf{DAD}^T$ with very good separation (right).*

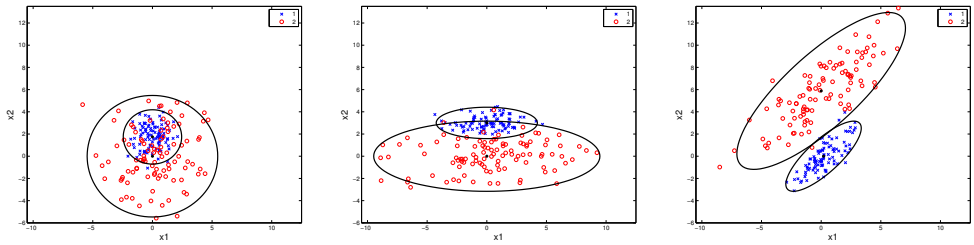


Figure 5.2: *Examples of simulated data with the volume changing across the mixture components: spherical model $\lambda_k\mathbf{I}$ with poor separation (left), diagonal model $\lambda_k\mathbf{A}$ with good separation (middle), and general model $\lambda_k\mathbf{DAD}^T$ with very good separation (right).*

the range $K = 1, \ldots, 5$ and the optimal number of mixture components was selected by using the Bayes factor (via the log marginal likelihoods). For these data sets, the used hyperparameters was as follows: $\boldsymbol{\mu}_0$ was equal to the mean of the data, the shrinkage $\kappa_n = 5$, the degree of freedom $\nu_0 = d + 2$, the scale matrix Λ_0 was equal to the covariance of the data, and the hyperparameter for the spherical models s_0^2 as the greatest eigenvalue of Λ_0.

5.2.2 Obtained results

Tables 5.2, 5.3 and 5.4 provide the obtained approximated log marginal likelihoods obtained by the PGMM and the proposed DPPM models, for, respectively, the equal (with equal clusters volumes) spherical data structure model ($\lambda\mathbf{I}$) and poorly separated mixture ($\varrho = 1$), the equal diagonal data structure model ($\lambda\mathbf{A}$) and good mixture separation ($\varrho = 3$), and the equal general data structure model ($\lambda\mathbf{DAD}^T$) and very good mixture separation ($\varrho = 4.5$). Tables 5.5, 5.6 and 5.7 provide the obtained approximated log marginal likelihoods obtained by the PGMM and the proposed DPPM models, for, respectively, the different (with different clusters volumes) spherical

data structure model ($\lambda_k\mathbf{I}$) and poorly separated mixture ($\varrho = 1$), the different diagonal data structure model ($\lambda_k\mathbf{A}$) with good mixture separation ($\varrho = 3$), and the different general data structure model ($\lambda_k\mathbf{DAD}^T$) with very good mixture separation ($\varrho = 4.5$).

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$\lambda\mathbf{I}$	2	-604.54	-633.88	-631.59	-635.07	-587.41	-595.63
$\lambda_k\mathbf{I}$	2	**-589.59**	-592.80	-589.88	-592.87	-593.26	-602.98
$\lambda\mathbf{A}$	2	-589.74	-591.67	-590.10	-593.04	-598.67	-599.75
$\lambda_k\mathbf{A}$	2	-591.65	-594.37	-592.46	-595.88	-607.01	-611.36
$\lambda\mathbf{DAD}^T$	2	-590.65	-592.20	**-589.65**	-596.29	-598.63	-607.74
$\lambda_k\mathbf{DAD}^T$	2	-591.77	-594.33	-594.89	-597.96	-594.49	-601.84

Table 5.2: *Log marginal likelihood values obtained by the proposed DPPM and PGMM for the generated data with $\lambda\mathbf{I}$ model structure and poorly separated mixture ($\varrho = 1$).*

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$\lambda\mathbf{I}$	2	-730.31	-771.39	-702.38	-703.90	-708.71	-840.49
$\lambda_k\mathbf{I}$	2	-702.89	-730.26	-702.30	-704.68	-708.43	-713.58
$\lambda\mathbf{A}$	2	**-679.76**	-704.40	**-680.03**	-683.13	-686.19	-691.93
$\lambda_k\mathbf{A}$	2	-685.33	-707.26	-688.69	-696.46	-703.68	-712.93
$\lambda\mathbf{DAD}^T$	2	-681.84	-693.44	-682.63	-688.39	-694.25	-717.26
$\lambda_k\mathbf{DAD}^T$	2	-693.70	-695.81	-684.63	-688.17	-694.02	-695.75

Table 5.3: *Log marginal likelihood values obtained by the proposed DPPM and the PGMM for the generated data with $\lambda\mathbf{A}$ model structure and well separated mixture ($\varrho = 3$).*

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$
$\lambda\mathbf{I}$	2	-762.16	-850.66	-747.29	-746.09	-744.63	-824.06
$\lambda_k\mathbf{I}$	2	-748.97	-809.46	-748.17	-751.08	-756.59	-766.26
$\lambda\mathbf{A}$	2	-746.05	-778.42	-746.32	-749.59	-753.64	-758.92
$\lambda_k\mathbf{A}$	2	-751.17	-781.31	-752.66	-761.02	-772.44	-780.34
$\lambda\mathbf{DAD}^T$	2	**-701.94**	-746.11	**-698.54**	-702.79	-707.83	-716.43
$\lambda_k\mathbf{DAD}^T$	2	-702.79	-748.36	-703.35	-708.77	-715.10	-722.25

Table 5.4: *Log marginal likelihood values obtained by the proposed DPPM and PGMM for the generated data with $\lambda\mathbf{DAD}^T$ model structure and very well separated mixture ($\varrho = 4.5$).*

From theses results, we can see that, the proposed DPPM, in all the situations (except for the first situation in Table 5.2) retrieves the actual model, with the actual number of clusters. We can also see that, except

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
$\lambda\mathbf{I}$	3	-843.50	-869.52	-825.68	-890.26	-906.44	-1316.40
$\lambda_k\mathbf{I}$	2	**-805.24**	-828.39	**-805.21**	-808.43	-811.43	-822.99
$\lambda\mathbf{A}$	2	-820.33	-823.55	-821.22	-825.58	-828.86	-838.82
$\lambda_k\mathbf{A}$	2	-808.32	-826.34	-808.46	-816.65	-824.20	-836.85
$\lambda\mathbf{DAD}^T$	2	-824.00	-823.72	-821.92	-830.44	-841.22	-852.78
$\lambda_k\mathbf{DAD}^T$	2	-821.29	-826.05	-803.96	-813.61	-819.66	-821.75

Table 5.5: *Log marginal likelihood values and estimated number of clusters for the generated data with $\lambda_k\mathbf{I}$ model structure and poorly separated mixture ($\varrho = 1$).*

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
$\lambda\mathbf{I}$	3	-927.01	-986.12	-938.65	-956.05	-1141.00	-1064.90
$\lambda_k\mathbf{I}$	3	-912.27	-944.87	-925.75	-911.31	-914.33	-918.99
$\lambda\mathbf{A}$	3	-899.00	-918.47	-906.59	-911.13	-917.18	-926.69
$\lambda_k\mathbf{A}$	2	**-883.05**	-921.44	-883.22	-897.99	-909.26	-928.90
$\lambda\mathbf{DAD}^T$	2	-903.43	-918.19	-902.23	-906.40	-914.35	-924.12
$\lambda_k\mathbf{DAD}^T$	2	-894.05	-920.65	**-876.62**	-886.86	-904.45	-919.45

Table 5.6: *Log marginal likelihood values obtained by the proposed DPPM and PGMM for the generated data with $\lambda_k\mathbf{A}$ model structure and well separated mixture ($\varrho = 3$).*

	DPPM		PGMM				
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$
$\lambda\mathbf{I}$	2	-984.33	-1077.20	-1021.60	-1012.30	-1021.00	-987.06
$\lambda_k\mathbf{I}$	3	-963.45	-1035.80	-972.45	-961.91	-967.64	-970.93
$\lambda\mathbf{A}$	2	-980.07	-1012.80	-980.92	-986.39	-992.05	-999.14
$\lambda_k\mathbf{A}$	2	-988.75	-1015.90	-991.21	-1007.00	-1023.70	-1041.40
$\lambda\mathbf{DAD}^T$	3	-931.42	-984.93	-939.63	-944.89	-952.35	-963.04
$\lambda_k\mathbf{DAD}^T$	2	**-921.90**	-987.39	**-921.99**	-930.61	-946.18	-956.35

Table 5.7: *Log marginal likelihood values obtained by the proposed DPPM and PGMM for the generated data with $\lambda_k\mathbf{DAD}^T$ model structure and very well separated mixture ($\varrho = 4.5$).*

for two situations, the selected DPPM model, has the highest log marginal likelihood value, compared to the PGMM. We also observe that the solutions provided by the proposed DPPM are, in some cases more parsimonious than those provided by the PGMM, and, in the other cases, the same as those provided by the PGMM. For example, in Table 5.2, which corresponds to data from poorly separated mixture, we can see that the proposed DPPM selects the spherical model $\lambda_k\mathbf{I}$, which is more parsimonious than the general model $\lambda\mathbf{A}$ selected by the PGMM, with a better misclassification error (see

Table 5.8). The same thing can be observed in Table 5.6 where the proposed DPPM selects the actual diagonal model $\lambda_k \mathbf{A}$, however the PGMM selects the general model $\lambda_k \mathbf{DAD}^T$, while the clusters are well separated ($\varrho = 3$).

Also in terms of misclassification error, as shown in Table 5.8, the proposed DPPM models, compared to the PGMM ones, provide partitions with the lower miscclassification error, for situations with poorly, well or very-well separated clusters, and for clusters with equal and different volumes (except for one situation).

PGMM	48 ± 8.05	9.5 ± 3.68	$\mathbf{1 \pm 0.80}$
DPPM	$\mathbf{40 \pm 4.66}$	$\mathbf{7 \pm 3.02}$	3 ± 0.97

Table 5.8: *Misclassification error rates obtained by the proposed DPPM and the PGMM approach. From left to right, the situations respectively shown in Table 5.2, 5.3, 5.4*

PGMM	23.5 ± 2.89	10.5 ± 2.44	2 ± 1.69
DPPM	$\mathbf{20.5 \pm 3.34}$	$\mathbf{7 \pm 3.73}$	$\mathbf{1.5 \pm 0.79}$

Table 5.9: *Misclassification error rates obtained by the proposed DPPM and the PGMM approach. From left to right, the situations respectively shown in Table 5.5, 5.6, 5.7*

On the other hand, for the DPMM models, from the log marginal likelihoods shown in Tables 5.2 to 5.7, we can see that the evidence of the selected model, compared to the majority of the other alternative is, according to Table 3.6, in general decisive. Indeed, it can be easily seen that the value $2 \log \mathrm{BF}_{12}$ of the Bayes Factor between the selected model, and the other models, is more than 10, which corresponds to a decisive evidence for the selected model. Also, if we consider the evidence of the selected model, against the more competitive one, one can see from Table 5.10 and Table 5.11, that, for the situation with very bad mixture separation, with clusters having the same volume, the evidence is not bad (0.3). However, for all the other situations, the optimal model is selected with an evidence going from an almost substantial evidence (a value of 1.7), to a strong and decisive evidence, especially for the models with different clusters volumes. We can also conclude that the models with different clusters volumes may work better in practice as highlighted by Celeux and Govaert (1995). Finally, Figure (5.3) shows the best estimated partitions for the data structures with equal volume across the mixture components shown in Fig. 5.1 and the posterior distribution over the number of clusters. One can see that for the case of clusters with equal volume, the diagonal family ($\lambda \mathbf{A}$) with well separated mixture ($\varrho = 3$) and the general family ($\lambda \mathbf{DAD}^T$) with very well separated mixture ($\varrho = 4.5$) data structure estimates a good number of clusters with

M_1 vs M_2	$\lambda_k\mathbf{I}$ vs $\lambda\mathbf{A}$	$\lambda\mathbf{A}$ vs $\lambda\mathbf{DAD}^T$	$\lambda\mathbf{DAD}^T$ vs $\lambda_k\mathbf{DAD}^T$
$2\log\mathrm{BF}$	0.30	4.16	1.70

Table 5.10: *Bayes factor values obtained by the proposed DPPM by comparing the selected model (denoted M_1) and the one more competitive for it (denoted M_2). From left to right, the situations respectively shown in Table 5.2, Table 5.3 and Table 5.4*

M_1 vs M_2	$\lambda_k\mathbf{I}$ vs $\lambda_k\mathbf{A}$	$\lambda_k\mathbf{A}$ vs $\lambda_k\mathbf{DAD}^T$	$\lambda_k\mathbf{DAD}^T$ vs $\lambda\mathbf{DAD}^T$
$2\log\mathrm{BF}$	6.16	22	19.04

Table 5.11: *Bayes factor values obtained by the proposed DPPM by comparing the selected model (denoted M_1) and the one more competitive for it (denoted M_2). From left to right, the situations respectively shown in Table 5.5, Table 5.6 and Table (6) 5.7*

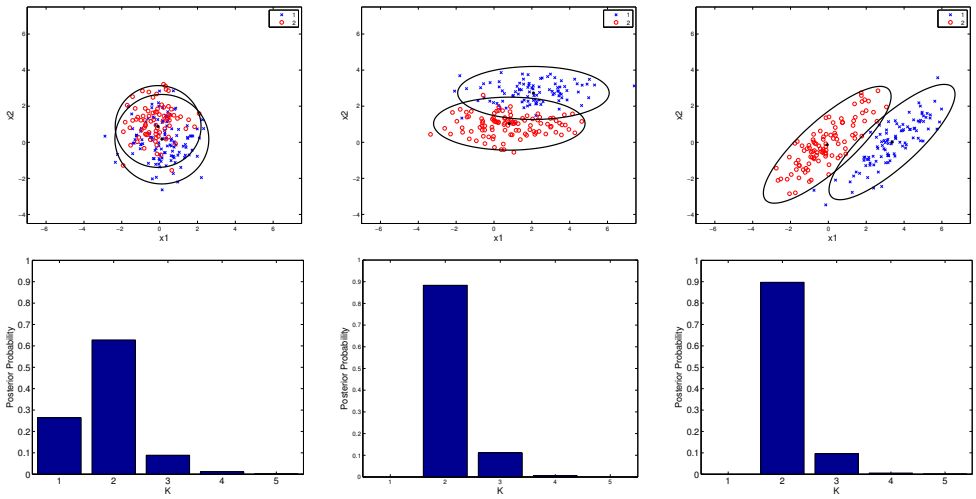


Figure 5.3: *Partitions obtained by the proposed DPPM for the data sets in Fig. 5.1.*

the actual model. However, the equal spherical data model structure ($\lambda\mathbf{I}$) estimates the $\lambda_k\mathbf{I}$ model, which is also a spherical model. Figure (5.4) shows the best estimated partitions for the data structures with different volume across the mixture components shown in Fig. 5.2 and the posterior distribution over the number of clusters. One can see that for all of different data structure models: different spherical $\lambda_k\mathbf{I}$, different diagonal $\lambda_k\mathbf{A}$ and different general $\lambda_k\mathbf{DAD}^T$, the proposed DPPM approach succeeded to estimate a good number of clusters equal to 2 with an actual cluster structure.

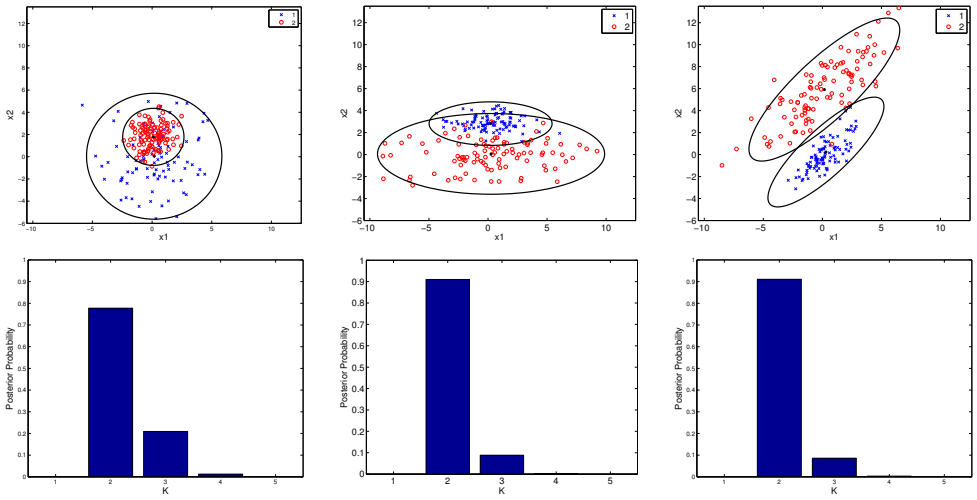


Figure 5.4: *Partitions obtained by the proposed DPPM for the data sets in Fig. 5.2.*

5.2.3 Stability with respect to the hyperparameters values

In order to illustrate the effect of the choice of the hyperparameters values of the mixture on the estimations, we considered two-class situations identical to those used in the parametric parsimonious mixture approach proposed in Bensmail et al. (1997). The data set consists in a sample of $n = 200$ observations from a two-component Gaussian mixture in $\mathbb{R}^2$ with the following parameters: $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (8, 8)^T$ and $\boldsymbol{\mu}_2 = (2, 2)^T$, and two spherical covariances with different volumes $\boldsymbol{\Sigma}_1 = 4\ \mathbf{I}_2$ and $\boldsymbol{\Sigma}_2 = \mathbf{I}_2$. In Figure (5.5) we can see a simulated data set from this experiment with the corresponding actual partition and density ellipses. In order to assess the stability of the models with respect to the values of the hyperparameters, we consider four situations with different hyperparameter values. These situations are as follows. The hyperparameters ν_0 and $\boldsymbol{\mu}_0$ are assumed to be the same for the four situations and their values are respectively $\nu_0 = d + 2 = 4$ (related to the number of degrees of freedom) and $\boldsymbol{\mu}_0$ equals the empirical mean vecotr of the data. We variate the two hyperparameters, κ_0 that controls the prior over the mean and s_0^2 that controls the covariance. The considered four situations are shown in Table 5.12. We consider and compare four mod-

Sit.	1	2	3	4
s_0^2	$\max(\text{eig}(\text{cov}(\mathbf{X})))$	$\max(\text{eig}(\text{cov}(\mathbf{X})))$	$4\ \max(\text{eig}(\text{cov}(\mathbf{X})))$	$\max(\text{eig}(\text{cov}(\mathbf{X})))/4$
κ_0	1	5	5	5

Table 5.12: *Four different situations the hyperparameters values.*

els corresponding to the spherical, diagonal and general family, which are

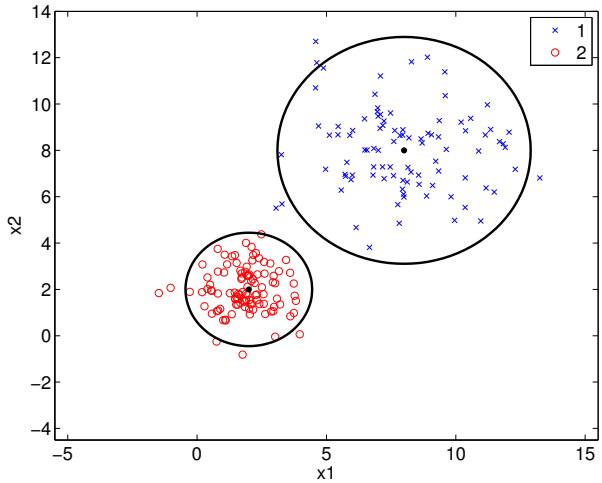


Figure 5.5: *A two-class data set simulated according to $\lambda_k\mathbf{I}$, and the actual partition.*

$\lambda\mathbf{I}$, $\lambda_k\mathbf{I}$, $\lambda_k\mathbf{A}$ and $\lambda_k\mathbf{DAD}^T$. Table 5.13 shows the obtained log marginal likelihood values for the four models for each of the situations of the hyperparameters. One can see that, for all the situations, the selected model is $\lambda_k\mathbf{I}$, that is the one that corresponds to the actual model, and has the correct number of clusters (two clusters). Also, it can be seen from Table

Model	$\lambda\mathbf{I}$		$\lambda_k\mathbf{I}$		$\lambda\mathbf{A}$		$\lambda_k\mathbf{DAD}^T$	
Sit.	$\hat{K}$	log ML	$\hat{K}$	log ML	$\hat{K}$	log ML	$\hat{K}$	log ML
1	2	-919.3150	2	**-865.9205**	3	-898.7853	3	-885.9710
2	3	-898.6422	2	**-860.1917**	2	-890.6766	2	-885.5094
3	2	-927.8240	2	**-884.6627**	2	-906.7430	2	-901.0774
4	2	-919.4910	2	**-861.0925**	2	-894.9835	2	-889.9267

Table 5.13: *Log marginal likelihood values for the proposed DPPM for 4 situations of hyperparameters values.*

5.14, that the Bayes factor values ($2\log\mathrm{BF}$), between the selected model, and the more competitive one, for each of the four situations, according to Table 3.6, corresponds to a decisive evidence of the selected model. These

Sit.	1	2	3	4
$2\log\mathrm{BF}$	40.10	50.63	32.82	57.66

Table 5.14: *Bayes factor values for the proposed DPPM computed from Table 5.13 by comparing the selected model (M_1, here in all cases $\lambda_k\mathbf{I}$), and the one more competitive for it (M_2, here in all cases $\lambda_k\mathbf{DAD}$).*

results confirm the stability of the DPPM with respect to the variation of

the hyparameters values. Figure 5.6 shows the best estimated partitions obtained by the proposed DPPM for the generated data. Note that, for the four situations, the estimated number of clusters equals 2 for all the situations, and the posterior mode of the distribution of the number of clusters is very close to 1.

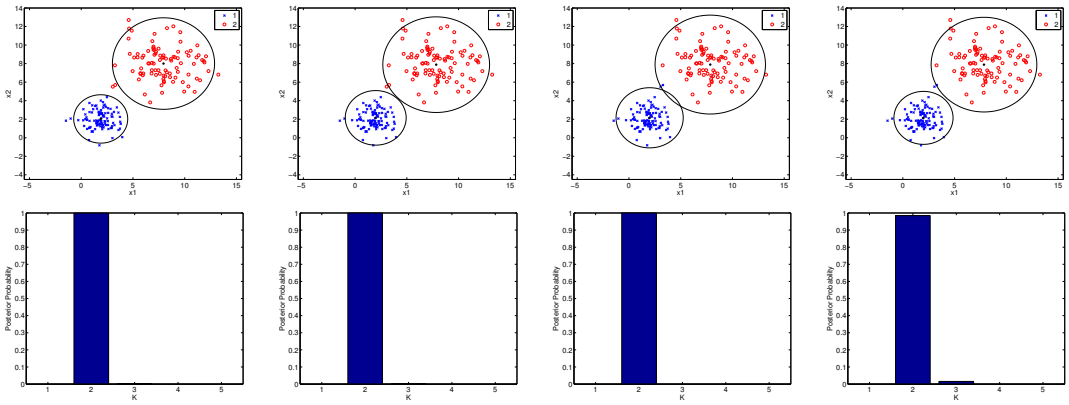


Figure 5.6: *Best estimated partitions obtained by the proposed $\lambda_k\mathbf{I}$ DPPM for the four situations of of hyperparameters values.*

5.3 Applications on benchmarks

To confirm the results previously obtained on simulated data, we have conducted several experiments freely available real data sets: Iris, Old Faithful Geyser, Crabs and Diabetes whose characteristics are summarized in Table 5.15. We compare the proposed DPPM models to the PGMM models.

Dataset	# data (n)	# dimensions (d)	True # clusters (K)
Old Faithful Geyser	272	2	Unknown
Crabs	200	5	2
Diabetes	145	3	3
Iris	150	4	3

Table 5.15: *Description of the used real data sets.*

5.3.1 Clustering of the Old Faithful Geyser data set

The Old Faithful geyser data set (Azzalini and Bowman, 1990) comprises $n = 272$ measurements of the eruption of the Old Faithful geyser at Yellowstone National Park in the USA. Each measurement is bi-dimensional ($d = 2$) and comprises the duration of the eruption and the time to the next

eruption, both in minutes. While the number of clusters for this data set is unknown, several clustering studies in the literature estimate at two, often interpreted as short and long eruptions.

We applied the proposed DPPM approach and the PGMM alternative to this data set (after standardization). For the PGMM, the value of K was varying from 1 to 6. Table 5.16 reports the log marginal likelihood values obtained by the PGMM and the proposed DPPM for the Faithful Geyser data set. One can see that the parsimonious DPPM models estimate 2 clus-

	DPPM		PGMM					
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
$\lambda\mathbf{I}$	2	-458.19	-834.75	-455.15	-457.56	-461.42	**-429.66**	-1665.00
$\lambda_k\mathbf{I}$	2	-451.11	-779.79	**-449.32**	-454.22	-460.30	-468.66	-475.63
$\lambda\mathbf{A}$	3	-424.23	-781.86	**-445.23**	-445.61	-445.63	-448.93	-453.44
$\lambda_k\mathbf{A}$	2	-446.22	-784.75	**-461.23**	-465.94	-473.55	-481.20	-489.71
$\lambda\mathbf{DAD}^T$	2	**-418.99**	-554.33	**-428.36**	-429.78	-433.36	-436.52	-440.86
$\lambda_k\mathbf{DAD}^T$	2	-434.50	-556.83	**-420.88**	-421.96	-422.65	-430.09	-434.36
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	2	-428.96	-780.80	-443.51	**-442.66**	-446.21	-449.40	-456.14
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	2	-421.49	-553.87	-434.37	**-433.77**	-439.60	-442.56	-447.88

Table 5.16: *Log marginal likelihood values for the Old Faithful Geyser data set.*

ters except one model, which is the diagonal model with equal volume $\lambda\mathbf{A}$ that estimates three clusters. For a number of clusters varying from 1 to 6, the parsimonious PGMM models estimate two clusters at three exceptions, including the spherical model $\lambda\mathbf{I}$ which overestimates the number of clusters (provides 5 clusters). However, the solution provided by the proposed DPMM for the spherical model $\lambda\mathbf{I}$ is more stable and estimates two clusters. It can also be seen that the best model with the highest value of the log marginal likelihood is the one provided by the proposed DPPM and corresponds to the general model $\lambda\mathbf{DAD}^T$ with equal volume and the same shape and orientation. On the other hand, it can also be noticed that, in terms of Bayes factors, the model $\lambda\mathbf{DAD}^T$ selected by the proposed DPMM has a decisive evidence compared to the other models, and a strong evidence (the value of $2\log\mathrm{BF}$ equals 5), compared to the most competitive one, which is in this case the model $\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$.

Figure 5.7 shows the the optimal partition and the posterior distribution for the number of clusters. One can namely observe that the likely partition is provided with a number of cluster with high posterior probability (more than 0.9).

Table 5.17 shows the mean computer running time, measured in seconds, for the Gibbs inference of each DPPM models.

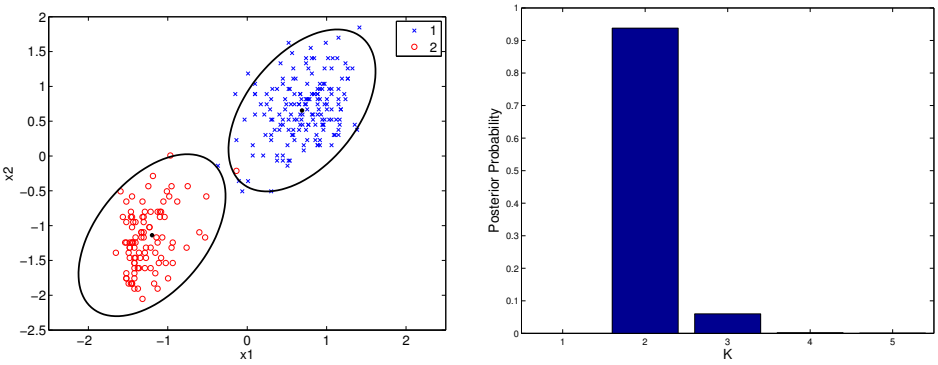


Figure 5.7: *Old Faithful Geyser data set (left), the optimal partition obtained by the DPPM model* $\lambda\mathbf{DAD}^T$ *(middle) and the empirical posterior distribution for the number of mixture components (right).*

Model	$\lambda\mathbf{I}$	$\lambda_k\mathbf{I}$	$\lambda\mathbf{A}$	$\lambda_k\mathbf{A}$	$\lambda\mathbf{DAD}^T$	$\lambda_k\mathbf{DAD}^T$	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$
CPU time (s)	953.86	785.36	999.91	964.86	901.44	717.28	1020	810.23

Table 5.17: *The DPPM Gibbs sampler mean CPU time (in seconds) for each parsimonious model on Old Faithful Geyser data set.*

5.3.2 Clustering of the Crabs data set

The Crabs data set comprises $n = 200$ observations describing $d = 6$ morphological measurements (Species, Frontal lip, Rearwidth, Length, Width Depth) on 50 crabs each of two colour forms and both sexes, of the species Leptograpsus variegatus collected at Fremantle, W. Australia Campbell and Mahon (1974). The Crabs are classified according to their sex ($K = 2$). We applied the proposed DPPM approach and the PGMM alternative to this data set (after PCA and standardization). For the PGMM the value of K was varying from 1 to 6. Table 5.18 reports the log marginal likelihood values obtained by the PGMM the proposed DPPM approaches for the Crabs data set. One can first see that the best solution corresponding to the best model

	DPPM		PGMM					
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$
$\lambda\mathbf{I}$	3	-550.75	-611.30	-615.73	**-556.05**	-860.95	-659.93	-778.21
$\lambda_k\mathbf{I}$	3	-555.91	-570.13	-549.06	-538.04	-542.31	-577.22	**-532.40**
$\lambda\mathbf{A}$	4	-537.81	-572.06	-539.17	-532.65	-535.20	-534.43	**-531.19**
$\lambda_k\mathbf{A}$	3	-543.97	-574.82	**-541.27**	-569.79	-590.48	-693.42	-678.95
$\lambda\mathbf{DAD}^T$	4	-526.87	-554.64	-540.87	**-512.78**	-525.19	-541.93	-576.27
$\lambda_k\mathbf{DAD}^T$	3	-517.58	-556.73	-541.88	**-515.93**	-530.02	-550.71	-595.38
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	4	-549.78	-573.80	-564.28	-541.67	-547.45	-547.13	**-526.79**
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	2	**-499.54**	-557.69	**-500.24**	-700.44	-929.24	-1180.10	-1436.60

Table 5.18: *Log marginal likelihood values for the Crabs data set.*

with the highest value of the log marginal likelihood is the one provided by the proposed DPPM and corresponds to the general model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ with different volume and orientation but equal shape. This model provides a partition with a number of clusters equal to the actual one $K = 2$. One can also see that the best solution for the PGMM approach is the one provided by the same model with a correctly estimated number of clusters. On the other hand, one can also see that for this Crabs data set, the proposed DPPM models estimate the number of clusters between 2 and 4. This may be related to the fact that, for the Crabs data set, the data, in addition their sex, are also described in terms of their specie and the data contains two species. This may therefore result in four subgroupings of the data in four clusters, each couple of them corresponding to two species, and the solution of four clusters may be plausible for this data set. However three PGMM models overestimate the number of clusters and provide solutions with 6 clusters. We can also observe that, in terms of Bayes factors, the model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ selected by the proposed DPMM for this data set, has a decisive evidence compared to all the other potential models. For example the value of $2 \log \mathrm{BF}$ for this selected model, against to the most competitive one, which is in this case the model $\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T$ equals 36.08 and corresponds to a decisive evidence of the selected model.

The good performance of the DPPM compared the PGMM is also confirmed in terms of Rand index and misclassification error rate values. The optimal partition obtained by the proposed DPPM with the parsimonious model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ is the best defined one and corresponds to the highest Rand index value of 0.8111 and the lowest error rate of 10.5 ± 1.98. However, the partition obtained by the PGMM has a Rand index of 0.8032 with an error rate of 11 ± 2.07.

Figure 5.8 shows the partition for Crabs data.

Figure 5.9 the optimal partition and the posterior distribution for the number of clusters. One can observe that the provided partition is quite precise and is provided with a number of clusters equal to the actual one, and with a posterior probability very close to 1.

Table 5.19 shows the mean computer running time, measured in seconds, for the Gibbs inference of each DPPM models.

Model	$\lambda \mathbf{I}$	$\lambda_k \mathbf{I}$	$\lambda \mathbf{A}$	$\lambda_k \mathbf{A}$	$\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$	$\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T$	$\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$
CPU time (s)	263.39	318.06	423.51	412.29	399.91	399.50	445.67	442.29

Table 5.19: *The DPPM Gibbs sampler mean CPU time (in seconds) for each parsimonious model on Crabs dataset.*

5.3.3 Clustering of the Diabetes data set

The Diabetes data set was described and analysed in (Reaven and Miller, 1979) consists of $n = 145$ subjects, describing $d = 3$ features: the area under

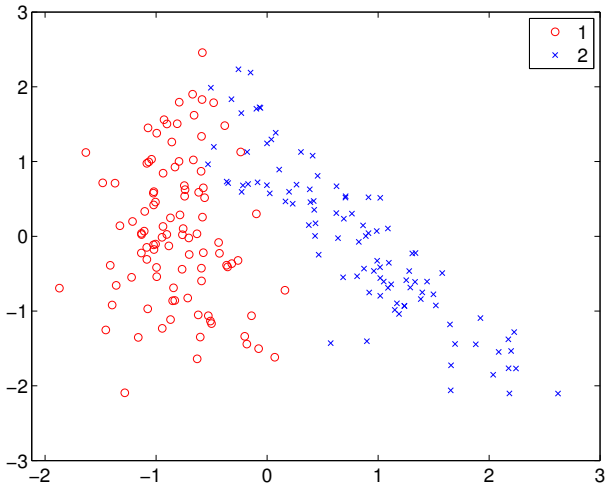


Figure 5.8: *Crabs data set in the two first principal axes and the actual partition.*

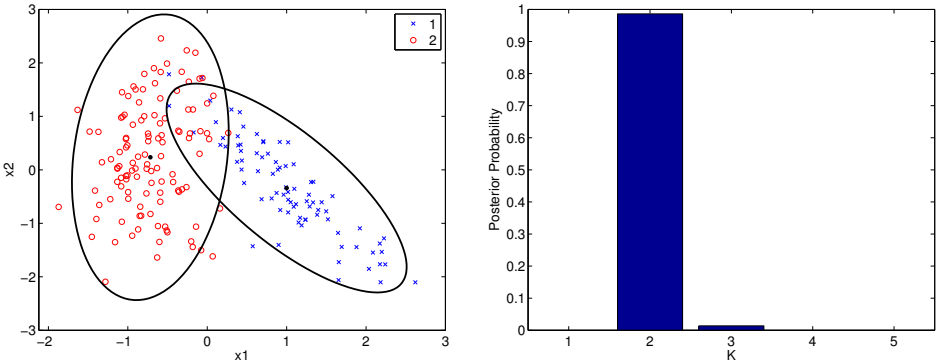


Figure 5.9: *The optimal partition obtained by the DPPM model* $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ *(middle) and the empirical posterior distribution for the number of mixture components (right).*

a plasma glucose curve (glucose area), the area under a plasma insulin curve (insulin area) and the steady-state plasma glucose response (SSPG). This data has $K = 3$ groups: the chemical diabetes, the overt diabetes and the normal (nondiabetic). We applied the proposed DPPM models and the alternative PGMM ones on this data set (the data was standardized). For the PGMM, the number of clusters was varying from 1 to 8.

Table 5.20 reports the log marginal likelihood values obtained by the two approaches for the Crabs data set. One can see that both the proposed DPPM and the PGMM estimate correctly the true number of clusters. However, the best model with the highest log marginal likelihood value is the one obtained by the proposed DPPM approach and corresponds to the parsimo-

nious model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ with the actual number of clusters ($K = 3$). Also,

	DPPM		PGMM							
Model	$\hat{K}$	log ML	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$	$K = 7$	$K = 8$
$\lambda \mathbf{I}$	4	-573.73	-735.80	-675.00	-487.65	-601.38	-453.77	-468.55	**-421.33**	-533.97
$\lambda_k \mathbf{I}$	7	-357.18	-632.18	-432.02	-412.91	-417.91	-398.02	-363.12	**-348.67**	-378.48
$\lambda \mathbf{A}$	8	-536.82	-635.70	-492.61	-488.55	-418.51	-391.05	-377.37	-370.47	**-365.56**
$\lambda_k \mathbf{A}$	6	-362.03	-638.69	-416.27	-372.71	**-358.45**	-381.68	-366.15	-385.73	-495.63
$\lambda \mathbf{D}\mathbf{A}\mathbf{D}^T$	7	-392.67	-430.63	-418.96	-412.70	**-375.37**	-390.06	-405.11	-426.92	-427.46
$\lambda_k \mathbf{D}\mathbf{A}\mathbf{D}^T$	5	-350.29	-432.85	-326.49	-343.69	**-325.46**	-355.90	-346.91	-330.11	-331.36
$\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	5	-338.41	-644.06	-427.66	-454.47	-383.53	-376.03	-356.09	-355.03	**-349.84**
$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	3	**-238.62**	-433.61	-263.49	**-248.85**	-273.31	-317.81	-440.67	-453.70	-526.52

Table 5.20: *Obtained marginal likelihood values for the Diabetes data set.*

the evidence of the model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ selected by the proposed DPMM for the Diabetes data set, compared to all the other models, is decisive. Indeed, in terms of Bayes factor comparison, the value of $2 \log \mathrm{BF}$ for this selected model, against to the most competitive one, which is in this case the model $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ is 111.86 and corresponds to a decisive evidence of the selected model. In terms of Rand index, the best defined partition is the one obtained by the proposed DPPM approach with the parsimonious model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$, which has the highest Rand index value of 0.8081 which indicates that the partition is well defined, with a misclassification error rate of 17.24 ± 2.47. However, the best PGMM partition $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ has a Rand index of 0.7615 with 22.06 ± 2.51 error rate.

Figure 5.10 shows the Diabetes data partition.

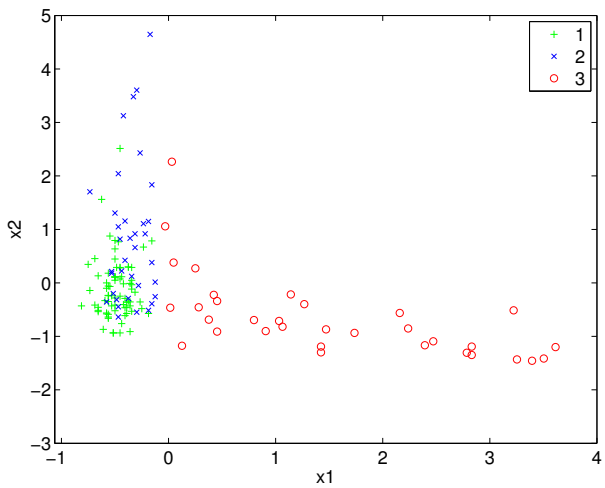


Figure 5.10: *Diabetes data set in the space of the components 1 (glucose area) and 3 (SSPG) and the actual partition.*

Figure (5.11) shows the optimal partition provided by the DPPM model

$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ and the distribution of the number of clusters K. We can observe that the partition is quite well defined (the misclassification rate in this case is 17.24 ± 2.47) and the posterior mode of the number of clusters equals the actual number of clusters ($K = 3$).

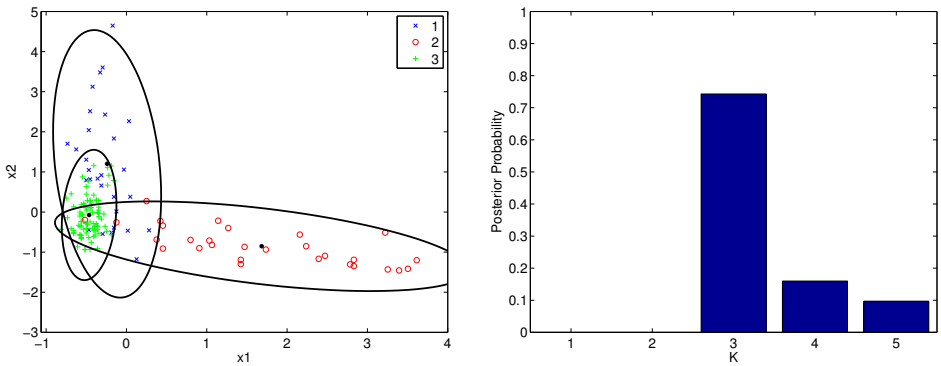


Figure 5.11: *The optimal partition obtained by the DPPM model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ (middle) and the empirical posterior distribution for the number of mixture components (right).*

Table 5.21 shows the mean computer running time, measured in seconds, for the Gibbs inference of each DPPM models.

Model	$\lambda \mathbf{I}$	$\lambda_k \mathbf{I}$	$\lambda \mathbf{A}$	$\lambda_k \mathbf{A}$	$\lambda \mathbf{D} \mathbf{A} \mathbf{D}^T$	$\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}^T$	$\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$	$\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$
CPU time (s)	1471.7	1335	1664	1386.8	1348.6	715.01	1635	1454.4

Table 5.21: *The DPPM Gibbs sampler mean CPU time (in seconds) for each parsimonious model on Diabetes data set.*

5.3.4 Clustering of the Iris data set

The first data set is Iris, well-known and was studied by Fisher Fisher (1936). It contains measurements for $n = 150$ samples of Iris flowers covering three Iris species (setosa, virginica and versicolor) ($K = 3$) with 50 samples for each specie. Four features were measured for each sample ($d = 4$): the length and the width of the sepals and petals, in centimetres. We applied PGMM models and the proposed DPPM models on this data set. For the PGMM models, the number of clusters K was tested in the range $[1; 8]$.

Table 5.22 reports the obtained log marginal likelihood values. We can see that the best solution is the one of the proposed DPPM and corresponds to the model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$, which has the highest log marginal likelihood value. One can also see that the other models provide partitions with two, three or four clusters and thus do not overestimate the number of clusters. However, the solution selected by the PGMM approach corresponds to a partition

with four clusters, and some of the PGMM models overestimate the number of clusters.

	DPPM		PGMM							
Model	$\hat{K}$	log ML	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$
$\lambda\mathbf{I}$	4	-415.68	-1124.9	-770.8	-455.6	-477.67	-431.22	-439.35	**-423.49**	-457.59
$\lambda_k\mathbf{I}$	3	-471.99	-913.47	-552.2	**-468.21**	-488.01	-507.8	-528.8	-549.62	-573.14
$\lambda\mathbf{A}$	3	-404.87	-761.44	-585.53	-561.65	-553.41	-546.97	-539.91	-535.37	**-530.96**
$\lambda_k\mathbf{A}$	3	-432.62	-765.19	**-623.89**	-643.07	-666.76	-688.16	-709.1	-736.19	-762.75
$\lambda\mathbf{DAD}^T$	4	-307.31	-398.85	-340.89	-307.77	**-286.96**	-291.7	-296.56	-300.37	-299.69
$\lambda_k\mathbf{DAD}^T$	2	-383.72	-401.61	-330.55	-297.50	**-279.15**	-282.83	-296.24	-304.37	-306.81
$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	4	-576.15	-1068.2	-761.71	-589.91	-529.52	-489.9	-465.37	**-444.84**	-457.86
$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	2	**-278.78**	-394.68	**-282.86**	-451.77	-676.18	-829.07	-992.04	-1227.2	-1372.8

Table 5.22: *Log marginal likelihood values for the Iris data set.*

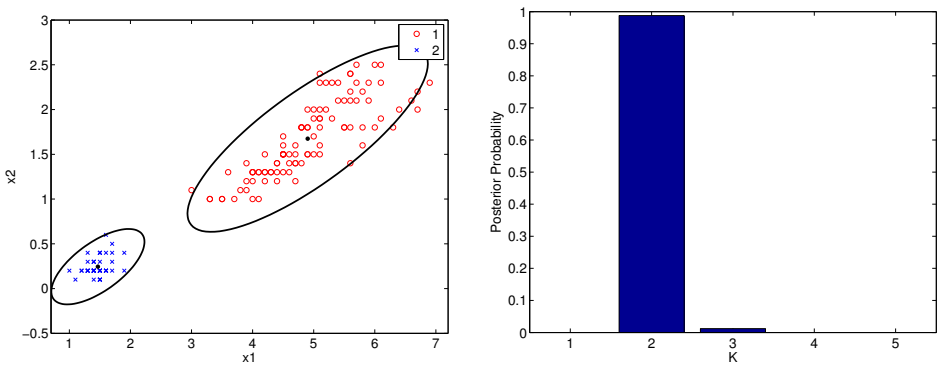


Figure 5.12: *The optimal partition obtained by the DPPM model* $\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$ *(middle) and the empirical posterior distribution for the number of mixture components (right).*

We also note that, the best partition found by the proposed DPPM, while in contains two clusters, is quite well defined, and has a Rand index of 0.7763.

Table 5.23 shows the mean computer running time, measured in seconds, for the Gibbs inference of each DPPM models.

Model	$\lambda\mathbf{I}$	$\lambda_k\mathbf{I}$	$\lambda\mathbf{A}$	$\lambda_k\mathbf{A}$	$\lambda\mathbf{DAD}^T$	$\lambda_k\mathbf{DAD}^T$	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$
CPU time (s)	144.04	261.34	342.48	352.81	293.91	382.0401	342.85	196.66

Table 5.23: *The DPPM Gibbs sampler mean CPU time (in seconds) for each parsimonious model on Iris data set.*

The evidence of the selected DPPM models, compared to the other ones, for the four real data sets, is significant. This can be easily seen in the tables showing the log marginal likelihood values. Consider the comparison

between the selected model, and the more competitive for it, for the four real data. As it can be seen in Table 5.24, which reports the values of $2\log \mathrm{BF}$ of the best model against the second best one, that the evidence of the selected model, according to Table 3.6 is strong for Old Faithful geyser data, and very decisive for Crabs, Diabetes and Iris data. Also, the model selection by the proposed DPMM for these latter three data sets, is made with a greater evidence, compared to the PGMM approach.

Data set	Old Faithful Geyser	Crabs	Diabetes	Iris
DPPM	$\lambda\mathbf{DAD}^T$ vs $\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$ vs $\lambda_k\mathbf{DAD}^T$	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$ vs $\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$ vs $\lambda\mathbf{DAD}^T$
$2\log \mathrm{BF}$	5	36.08	199.58	57.06
PGMM	$\lambda_k\mathbf{DAD}^T$ vs $\lambda\mathbf{DAD}^T$	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$ vs $\lambda\mathbf{DAD}^T$	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$ vs $\lambda_k\mathbf{DAD}^T$	$\lambda_k\mathbf{DAD}^T$ vs $\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$
$2\log \mathrm{BF}$	14.96	25.08	153.22	7.42

Table 5.24: *Bayes factor values for the selected model against the more competitive for it, obtained by the PGMM and the proposed DPPM for the real data sets.*

5.4 Scaled application on real-world bioacoustic data

In this section, we will apply the DPPM models on a further real dataset in the framework of a challenging problem of humpback whale song decomposition. The objective is the unsupervised structuration of these bioacoustic data. Humpback whale songs are long cyclical sequences produced by males during the reproduction season which follows their migration from high-latitude to low-latitude waters. Singers of one geographical population share parts of the same song. This leads to the idea of dialect (Helweg et al., 1998). Different hypotheses of these songs were emitted (Baker and Herman, 1984; Frankel et al., 1995; Garland et al., 2011; Medrano et al., 1994; Mercado and Kuh, 1998), even as used as sonar (Au et al., 2001; Frazer and Mercado, 2000).

Data description

The data consist in whale song signals in the framework of unsupervised analysis of bioacoustic data. This humpback whale song recording has been produced at few meters distance from the whale in La Reunion - Indian Ocean, by the "Darewin" regroup in 2013, at a Frequency Sample of 44.1kHz, 32 bits, mono, wav format.

They consist of MFFC features of 8.6 minutes that have been extracted using Spro 5.0, with pre-emphasis: 0.95, hamming window, fft on 1024 points (nearly 23ms), frameshift 10 ms, 24 Mel channels, 12 MFCC coefficients and energy and their delta and acceleration, CMS (mean normalisation)

and variance normalization, for a total of 39 dimensions as detailed in the SABIOD NIPS challenge : http://sabiod.univ-tln.fr/nips4b/challenge2.html where the signal and the features are available.

A spectrum of this whale of around 20 seconds of the given song can be seen in Figure 5.13. The data comprises 51336 observations with 39 features.

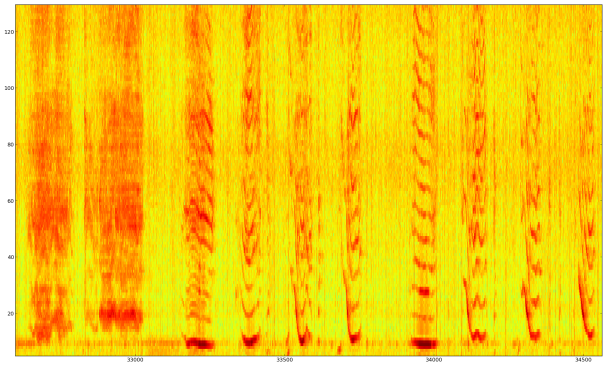


Figure 5.13: *Spectrum of around* 20 *seconds of the given song of Humpback Whale (start from about 5'40 to 6'). Ordinata from 0 to 22.05 kHz, over 512 bins (FFT on 1024 bins), frameshift of 10 ms.*

A dimension reduction pretreatment with a PCA technique was made. We therefore choose to retain 13 features of the data, since it was sufficient to capture more then 95% of the cumulative percentage of the variance.

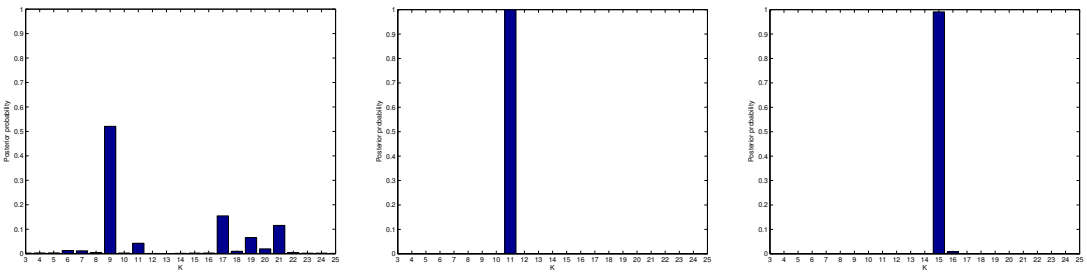
The analysis of such complex signals that aims at discovering the call units (which can be considered as a kind of whale alphabet), can be seen as a problem of unsupervised call units classification as in Pace et al. (2010). Another analysis of the humpback whale song by clustering approach can be found in Picot et al. (2008). The authors in Picot et al. (2008) implemented a segmentation algorithm based on Payne's principle to extract sound units of a whale song. In their application, six song units (pattern intonations) were found. We therefore reformulate the problem of whale song decomposition as an unsupervised data classification problem. Contrary to the approach used in Pace et al. (2010), in which the number of states (call units in this case) has been fixed manually, or Picot et al. (2008) where the unsupervised algorithm K-means was performed for automatic classification and then automatically define the optimal number of classes by maximizing the Davies Bouldin criterion. here, we first apply the proposed DPPM models to learn the complex bioacoustic data, to find the classes (states) of the whale song, and automatically infer the number of classes (states) from the data.

Unsupervised structuration of whale song data with the proposed DPPM models

We applied our proposed DPPM approach, into the challenging problem of Whale song decomposition NIPS4B challenge (Bartcus et al., 2013).

The Gibbs sampling runs 10 times with 4000 samplers and a burn-in period equal to 10%, by selecting the one with the highest MAP. Covering the three families, from the simplest one, which are the spherical models ($\lambda\mathbf{I}$ and $\lambda_k\mathbf{I}$), the diagonal models ($\lambda\mathbf{A}$ and $\lambda_k\mathbf{A}$), to the more complex general models ($\lambda\mathbf{DAD}^T$, $\lambda_k\mathbf{DAD}^T$ and $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$) are applied in this application.

In Figure 5.14 we show the posterior distributions of the numbers of components provided by the Gibbs sampler for the spherical model $\lambda\mathbf{I}$, the diagonal model $\lambda_k\mathbf{A}$ and the general model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$. We can see that model $\lambda\mathbf{I}$ retrieves 9 clusters, the model $\lambda_k\mathbf{A}$ retrieves 11 clusters and model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ retrieves 15 clusters.



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

Figure 5.14: *Posterior distribution of the number of components obtained by the proposed DPPM approach, for the whale song data.*

Because of the length of 8.6 minutes of the signal, for a more detailed information, we show separate parts of 15 seconds of the whole signal of the humpback whale. Some examples of the humpback whale song with 15 seconds duration each are presented. First, in Figure 5.15, we show two different signals with top, the signal starting at 45 seconds and it's corresponding partition obtained by the proposed DPM model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ (general), and bottom those for the part of the signal starting at 60 seconds. Then in Figure 5.16, we show the two different signals with top, the signal starting at 240 seconds and it's corresponding partition obtained by the proposed DPM model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ (general), and bottom those for the part of the signal starting at 255 seconds. Finally, in Figure 5.17 we show the two different signals with top, the signal starting at 280 seconds and it's corresponding partition obtained by the proposed DPM model $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ (general), and bottom those for the part of the signal starting at 295 seconds.

Next, we illustrate the obtained results for the two proposed DPPM

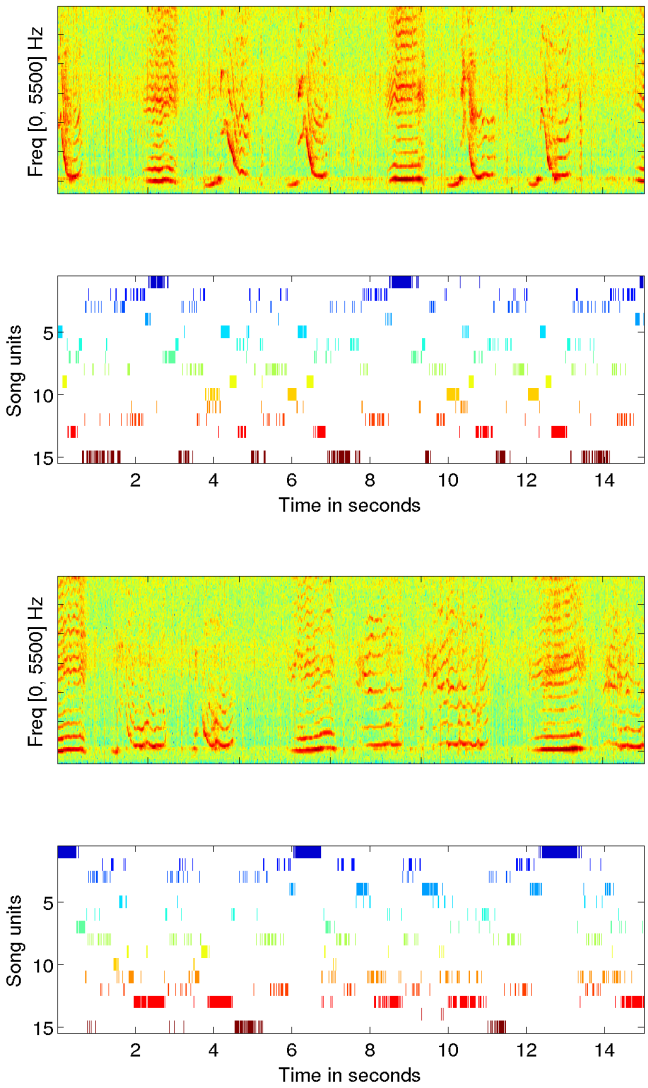


Figure 5.15: *Obtained song units by applying or DPM model with the parametrization* $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ *(general) to two different signals with top: the spectrogram of the part of the signal starting at 45 seconds and it's corresponding partition, and bottom those for the part of signal starting at 60 seconds.*

models, that corresponds to the parsimonious spherical model $\lambda \mathbf{I}$ with equal cluster volumes and the parsimonious diagonal model $\lambda_k \mathbf{A}$ with different cluster volumes. As for the general model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$, we show separate parts of 15 seconds duration of the whole signal of the humpback whale song in order to visualize the signal in a more detail.

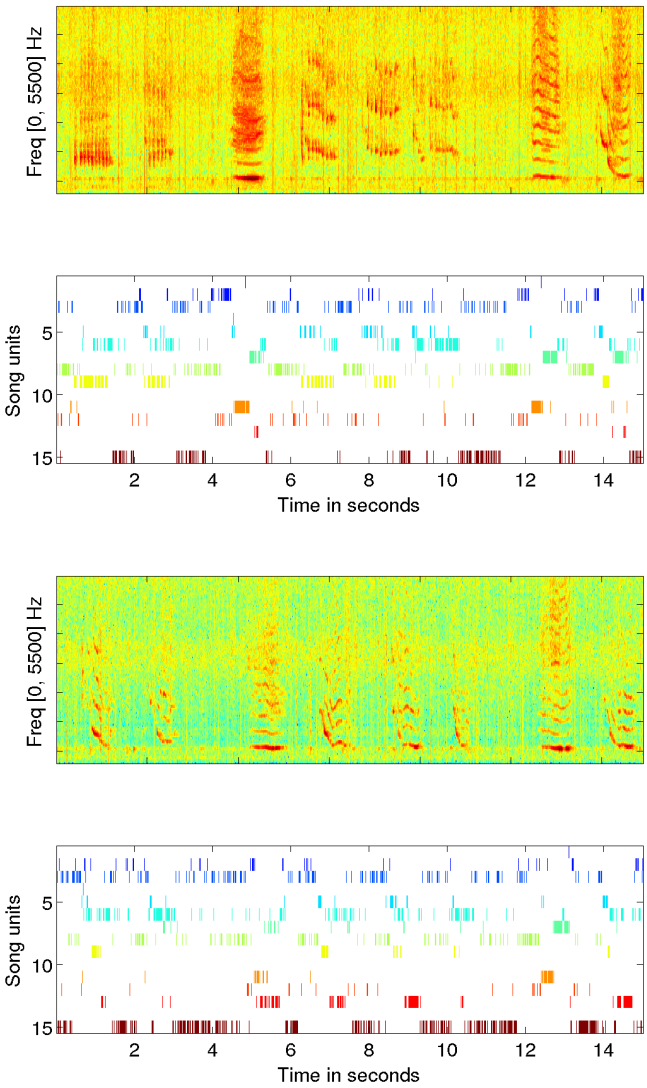


Figure 5.16: *Obtained song units by applying or DPM model with the parametrization* $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ *(general) to two different signals with top: the spectrogram of the part of the signal starting at 240 seconds and it's corresponding partition, and bottom those for the part of signal starting at 255 seconds.*

First, in Figure 5.18, we show two different signals with top, the signal starting at 45 seconds and it's corresponding partition obtained by the proposed DPPM model $\lambda \mathbf{I}$ (spherical), and bottom those for the part of the signal starting at 60 seconds.

Figure 5.19, shows two different signals with top, the signal starting

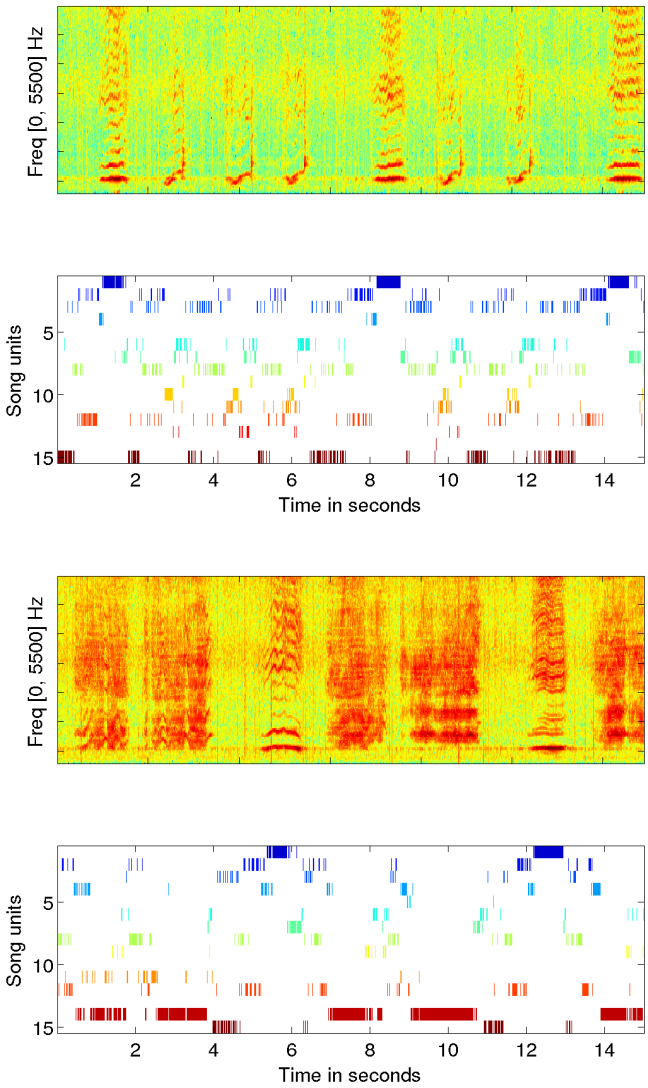


Figure 5.17: *Obtained song units by applying or DPM model with the parametrization* $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ *(general) to two different signals with top: the spectrogram of the part of the signal starting at 280 seconds and it's corresponding partition, and bottom those for the part of signal starting at 295 seconds.*

at 240 seconds and it's corresponding partition obtained by the proposed DPPM model $\lambda\mathbf{I}$ (spherical), and bottom those for the part of the signal starting at 255 seconds. Finally, Figure 5.20, shows two different signals with top, the signal starting at 280 seconds and it's corresponding partition obtained by the proposed DPM model $\lambda\mathbf{I}$ (spherical), and bottom those for

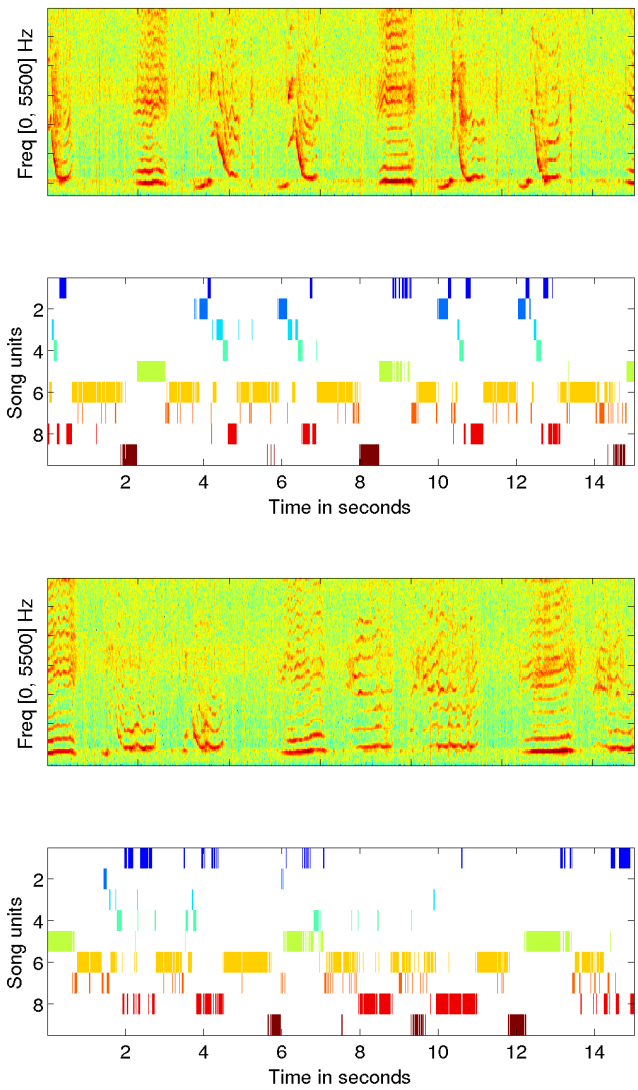


Figure 5.18: *Obtained song units by applying or DPPM model with the parametrization $\lambda\mathbf{I}$ (spherical) to two different signals with top: the spectrogram of the part of the signal starting at 45 seconds and it's corresponding partition, and bottom those for the part of signal starting at 60 seconds.*

the part of the signal starting at 295 seconds.

The spherical $\lambda\mathbf{I}$ model fit well the whale song data set with 9 song units. In this situation, it is noticed that the sixth state represents the silence, that can be filled with state 7 and 8. The state 4 is a very noisy and broad sound.

We also show the several parts of 15 seconds duration each, obtained by the proposed DPPM model $\lambda_k\mathbf{A}$ (diagonal). Figure 5.21, shows the signal

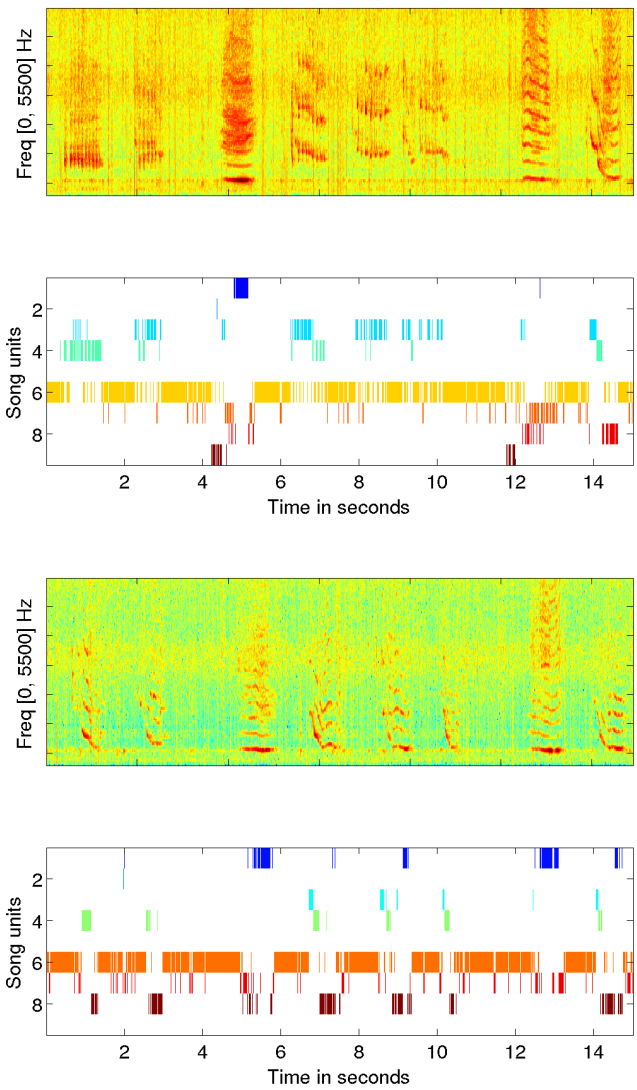


Figure 5.19: *Obtained song units by applying or DPPM model with the parametrization* $\lambda\mathbf{I}$ *(spherical) to two different signals with top: the spectrogram of the part of the signal starting at 240 seconds and it's corresponding partition, and bottom those for the part of signal starting at 255 seconds.*

starting with 45 seconds and it's corresponding obtained partition (top), and those for the part of the signal starting with 60 seconds (bottom). Figure 5.22, shows the signal starting with 240 seconds and it's corresponding obtained partition (top), and those for the part of the signal starting with 255 seconds (bottom). Figure 5.22, shows the signal starting with 280 seconds and it's corresponding obtained partition (top), and those for the

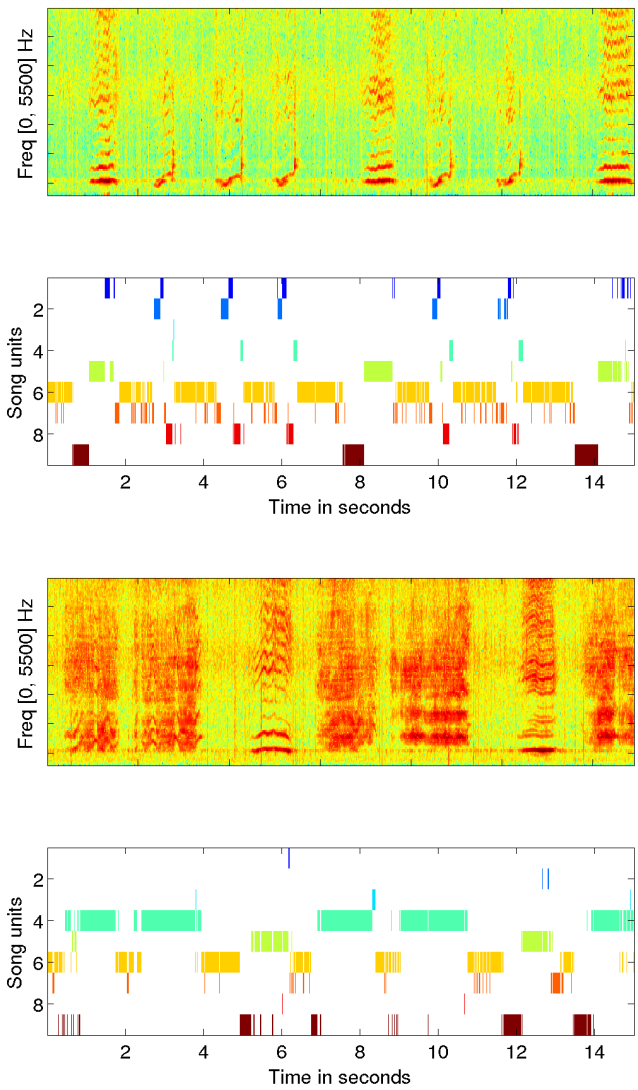


Figure 5.20: *Obtained song units by applying or DPPM model with the parametrization* $\lambda\mathbf{I}$ *(spherical) to two different signals with top: the spectrogram of the part of the signal starting at 280 seconds and it's corresponding partition, and bottom those for the part of signal starting at 295 seconds.*

part of the signal starting with 295 seconds (bottom).

The DPPM diagonal model, with different cluster volumes, that corresponds to the covariance matrix decomposition $\lambda_k\mathbf{A}$ fit well the data with 11 song units. It can clearly be seen that the state 9 is the silence. State 1, 2, 8 and 11 is the up and down sweeps. The seventh state is also the silence that generally ends the ninth state. The state 4 is a very noisy and broad

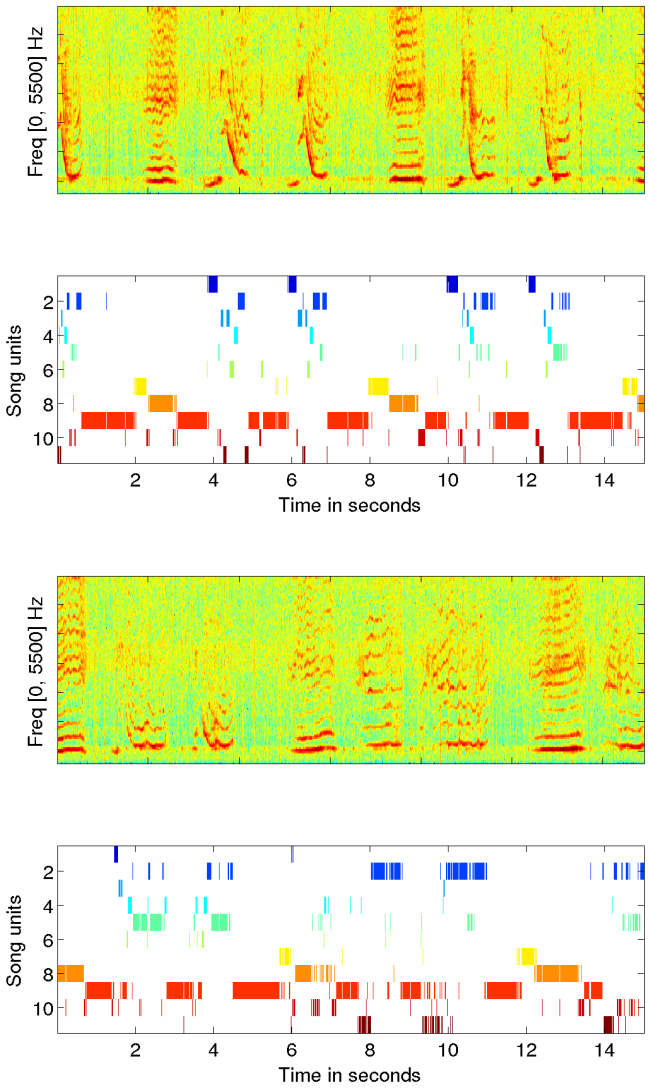


Figure 5.21: *Obtained song units by applying or DPPM model with the parametrization $\lambda_k \mathbf{A}$ (diagonal) to two different signals with top: the spectrogram of the part of the signal starting at 45 seconds and it's corresponding partition, and bottom those for the part of signal starting at 60 seconds.*

sound. The obtaining results highlighted the interest of using parsimonious Bayesian non-parametric modeling such that, even if they are not derived for sequential data.

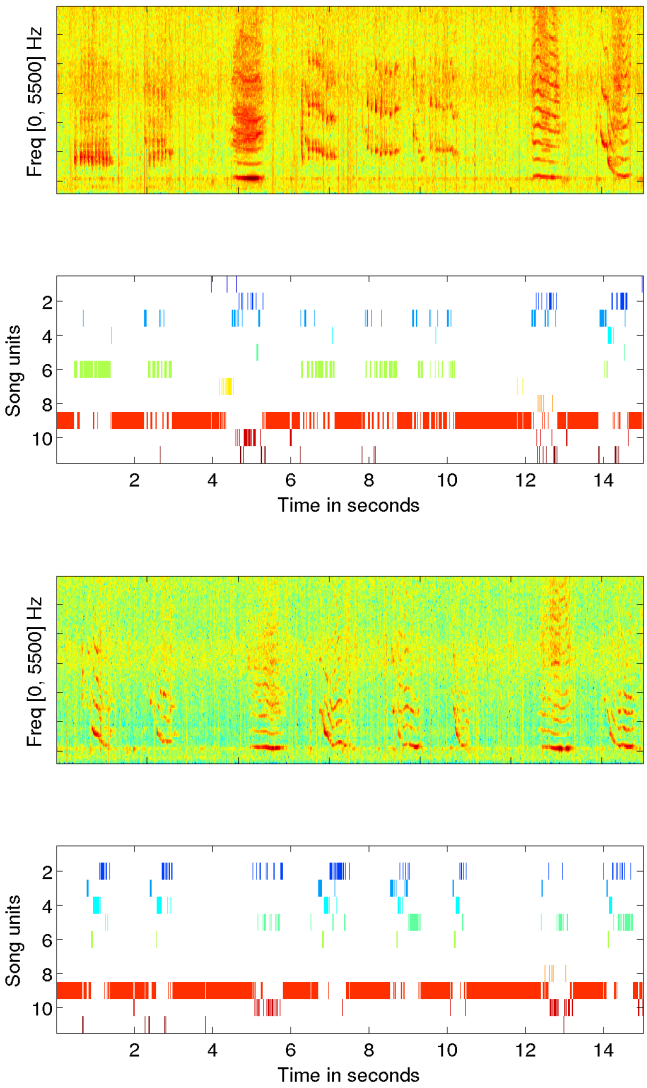


Figure 5.22: *Obtained song units by applying or DPPM model with the parametrization $\lambda_k \mathbf{A}$ (diagonal) to two different signals with top: the spectrogram of the part of the signal starting at 240 seconds and it's corresponding partition, and bottom those for the part of signal starting at 255 seconds.*

5.5 Conclusion

This chapter was dedicated to experiments of simulated and real-world data sets. It highlighted that the proposed DPPM represent a good non-parametric alternative of the model selection problem to the standard para-

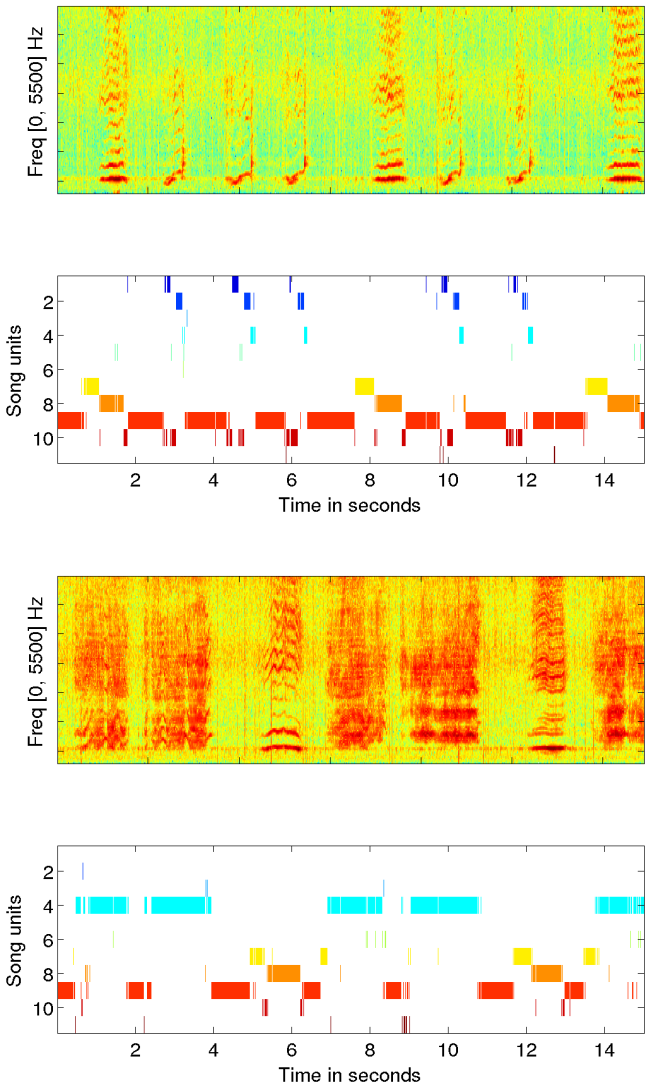


Figure 5.23: *Obtained song units by applying or DPPM model with the parametrization* $\lambda_k \mathbf{A}$ *(diagonal) to two different signals with top: the spectrogram of the part of the signal starting at 280 seconds and it's corresponding partition, and bottom those for the part of signal starting at 295 seconds.*

metric Bayesian and non-Bayesian finite mixtures. They simultaneously and accurately estimate accurate partitions with the optimal number of clusters inferred from the data. The optimal data structure is selected with using the Bayes Factor. The obtained results show the interest of using the Bayesian parsimonious clustering models and the potential benefit of using them in practical applications.

We applied the models on the challenging problem of humpback whale song decomposition. Despite the fact that the dataset are by nature sequential, and DPPMs models assume an exchangeability property, the models arrive to fit quiet satisfying partition of the data. This application opens a perspective on the extension of the previously discussed DPPMs models, from the i.i.d case to sequential data. Hence this may provide a good perspective for further integrating the parsimonious DPM models into a Markovian framework.

In the next chapter we investigate the Bayesian non-parametric extension of the standard Markovian framework proposed by (Beal et al., 2002; Teh et al., 2006). These Bayesian non-parametric HMM model, being tailored to sequential data, opens great perspective for future extensions of the DPPM models.

- Chapter 6 -

Bayesian non-parametric Markovian perspectives

Contents

6.1 Introduction

In Chapter 4, we proposed an extension for the BNP modeling for GMMs to the parsimonious BNP modeling. In Section 5.4, we applied the proposed approach on the complex bioacoustic signal. The obtained results fit the data despite the fact that the data is by nature sequential. Hidden Markov Models (HMM) (Rabiner, 1989) being one of the most successful models for modeling sequential data will open a Markovian perspective for the BNP modeling of the HMM.

In this chapter, we rely on the Hierarchical Dirichlet Process for Hidden Markov Models (HDP-HMM) proposed in (Beal et al., 2002; Teh et al., 2006) to investigate the challenging problem of unsupervised learning from bioacoustic data as in (Bartcus et al., 2015). Recall that this problem of fully unsupervised humpback whale song decomposition, as previously described in Section 5.4, consists in simultaneously finding the structure of hidden whale song units, and automatically inferring the unknown number of the hidden units from the Mel Frequency Cepstral Coefficients (MFCC) of bioacoustic signals. The experimental results shows very good performances of the proposed Bayesian non-parametric approach and opens new insights for unsupervised analysis of such bioacoustic signals. We use Markov-Chain Monte Carlo (MCMC) sampling techniques, particularly the Gibbs sampler, as in Fox (2009); Fox et al. (2008); Teh et al. (2006), to infer the HDP-HMM from the bioacoustic data.

This chapter is organized as follows. Section 6.2 describes the model and the inference technique using Gibbs sampling. The Section 6.3 is dedicated to it's application to the unsupervised decomposition of bioacoustic signals.

6.2 Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM)

Previously we saw that for the BNP modeling approach for the GMMs, Dirichlet Process prior were sufficient to extend the GMM to the infinite GMM case. However, for the HMM, where the transitions of states take independent priors, that is, there is no coupling across transitions between the different states (Beal et al., 2002). The Dirichlet Process (Ferguson, 1973) is not sufficient to extend HMM to an infinite state space model. The Hierarchical Dirichlet Process (HDP) prior (Teh et al., 2006) over the transition matrix (Beal et al., 2002) tackle this issue and extends the HMM to the infinite state space model.

Hierarchical Dirichlet Process (HDP)

Recalling the Dirichlet Process (DP) (Ferguson, 1973), that is a prior distribution over distributions, denoted as $\mathrm{DP}(\alpha, G_0)$ with two parameters, the scaling parameter α and the base measure G_0. The DP extends the finite modeling to the infinite modeling. However DP is not sufficient to extend HMM to an infinite state space model. In this section we refer on observations organized into groups, where it is supposed j refers to the groups and i the observations of each group. Thus we assume $\mathbf{x}_j = (\mathbf{x}_{j1}, \mathbf{x}_{j2}, \ldots, \mathbf{x}_{jn})$ denotes all exchangeable observations of group j. The groups observations $\mathbf{x}_1, \mathbf{x}_2 \ldots$ are in turn exchangeable. So, in this situation, when the data has a related but different generative process, the Hierarchical Dirichlet Process (HDP) prior is used to extend the HMM to an infinite state space HDP-HMM (Teh et al., 2006). A HDP assumes that the random measures

$$G_j|\alpha, G_0 \sim \mathrm{DP}(\alpha, G_0), \forall k = 1, \ldots K, \tag{6.1}$$

are itself distributed according to the DP with the hyperparameter α and the base measure G_0 that is in turn distributed by the DP with the hyperparameters γ and base distribution H.

$$G_0|\gamma, H \sim \mathrm{DP}(\gamma, H). \tag{6.2}$$

A HDP can be used as a prior distribution for factors of the grouped data. Suppose for each j, $\boldsymbol{\theta}_{j1}, \boldsymbol{\theta}_{j2}, \ldots, \boldsymbol{\theta}_{jn}$ be i.i.d random variables distributed by the G_j. Then, $\boldsymbol{\theta}_{ji}$ will be the parameter corresponding to each single observation $\mathbf{x}_{ji}$. So, the following completes the hierarchical Dirichlet process:

$$\begin{array}{rcl} \boldsymbol{\theta}_{ji}|G_j & \sim & G_j, \\ \mathbf{x}_{ji}|\boldsymbol{\theta}_{ji} & \sim & F(\mathbf{x}_{ji}|\boldsymbol{\theta}_{ji}). \end{array} \tag{6.3}$$

As a result the probabilistic graphical model for the hierarchical Dirichlet Process mixture model can be illustrated as follows:

Chinese Restaurant Franchise (CRF)

The Chinese Restaurant Process plays a great role in the representation of the Dirichlet Process, by giving a metophor to the existence of a restaurant with possible infinite tables (clusters) that customers (the observations) are siting in that restaurant. An alternative of such a representation for the Hierarchical Dirichlet Process can be described by the Chinese Restaurant Franchise process by extending the CRP to a multiple restaurants that shares a set of dishes.

The Chinese Restaurant Franchise (CRF) gives a representation for the Hierarchical Dirichlet Process (HDP) by extending the Chinese Restaurant Process (CRP) (Pitman, 1995; Samuel and Blei, 2012; Wood et al., 2006) to

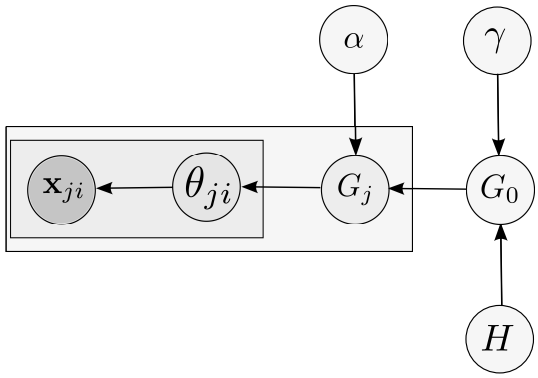


Figure 6.1: *Probabilistic Graphical Model for Hierarchical Dirichlet Process Mixture Model.*

a set of (J) restaurants rather than a single restaurant. Suppose a patron of Chinese Restaurant creates many restaurants, strongly linked to each other, by a franchise wide menu, having dishes common to all restaurants. As a result, J restaurants are created (groups) with a possibility to extend each restaurant to an infinite number of tables (states) at witch the customers (observations) sit. Each customer goes to his specified restaurant j, where each table of this restaurant has a dish that shares between the customers that sit at that specific table. However, multiple tables of different existing restaurants can serve the same dish. Figure 6.2 represents one such Chinese Restaurant Franchise Process for 2 restaurants. One can see the customers $\mathbf{x}_{ji}$ enters the restaurant j and takes the place of a table t_{ji}. Each table has a specific dish k_{jt} that can be also common for different restaurants.

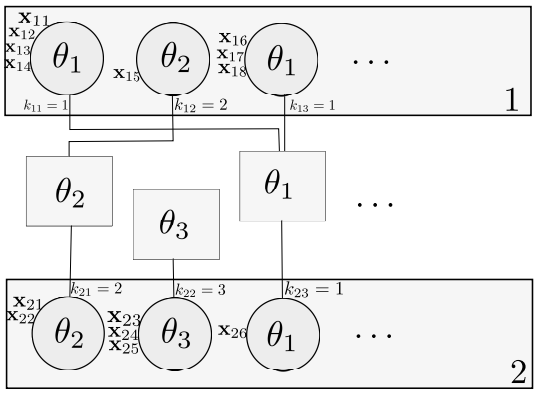


Figure 6.2: *Representation of a Chinese Restaurant Franchise with* 2 *restaurants. The clients* $\mathbf{x}_{ji}$ *are entering the* j*th restaurant* ($j = \{1, 2\}$)*, sit at table* t_{ji} *and chose the dish* k_{jt}*.*

The generative process of the Chinese Restaurant Franchise can be for-

mulated as follows. For each table a dish is assigned with $k_{jt}|\beta \sim \beta$, where β is the rating of the dish served at the specific restaurant j. The table assignment of the jth restaurant for the ith customer is then drawn. Finally the observations, $\mathbf{x}_{ji}$, or the customers i that enters the restaurant j are generated by a distribution $F(\boldsymbol{\theta}_{k_{jt_{ji}}})$. The generative process for CRF is given by the following:

$$\begin{aligned} k_{jt}|\beta \sim & \beta \\ t_{ji}|\tilde{\boldsymbol{\pi}}_j \sim & \tilde{\boldsymbol{\pi}}_j \\ \mathbf{x}_{ji}|\{\boldsymbol{\theta}_k\}_{k=1}^{\infty}, \{k_{jt}\}_{t=1}^{\infty}, t_{ji} \sim & F(\boldsymbol{\theta}_{k_{jt_{ji}}}) \end{aligned} \tag{6.4}$$

A probabilistic graphical model of such a process can be seen in the Figure 6.3.

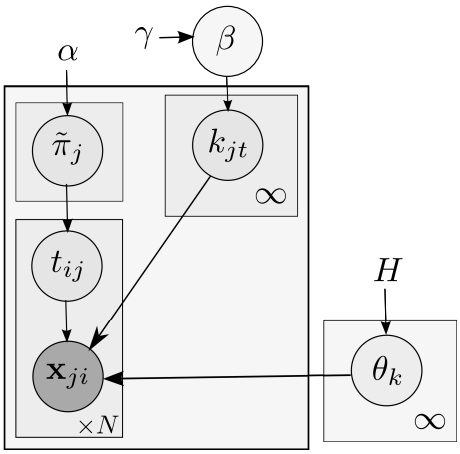


Figure 6.3: *Probabilistic graphical representation of the Chinese Restaurant Franchise (CRF).*

More details for derivation and inference of the Chinese Restaurant Franchise (CRF) and the use of it in the Hierarchical Dirichlet Process could be found in Teh and Jordan (2010); Teh et al. (2006) and Fox (2009); Fox et al. (2008).

An HDP-HMM representation as an Infinite Hidden Markov Model (IHMM)

The idea of the infinite mixture models for sequential data appears naturally after great performances with the i.i.d data, where the number of clusters were chosen in an automatic way instead of using some cross validation task. Due to the fact that the HMMs are one of the most popular and successful models in statistics and machine learning for modeling sequential data, it was meant to be developed to the infinite Hidden Markov Model.

It was shown that, by using the Dirichlet processes theory, more exactly the Hierarchical Dirichlet Process, it was possible to extending the Hidden Markov models into the infinite countable hidden number of states (Beal et al., 2002; Fox, 2009; Fox et al., 2008; Teh and Jordan, 2010; Teh et al., 2006; Van Gael et al., 2008).

Hierarchical Bayesian formulation gives the possibility to have distributions over hyper-parameters by making the models more flexible. The coupling between transition matrix allows a higher level to DP prior over the parameters.

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{Dir}(\gamma/K, \ldots, \gamma/K) \\ \boldsymbol{\pi}_k &\sim \mathrm{Dir}(\alpha\boldsymbol{\beta}) \end{aligned} \tag{6.5}$$

$\boldsymbol{\pi}_k$ being the transition matrix for the specific group k and $\boldsymbol{\beta}$ the prior hyperparameter.

Let G_k describes both, the transition matrix $\boldsymbol{\pi}_k$ and the emission parameters $\boldsymbol{\theta}_k$, the infinite HMM can be described by the following generative process:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\pi}_k|\alpha, \boldsymbol{\beta} &\sim \mathrm{DP}(\alpha, \boldsymbol{\beta}) \\ z_t|z_{t-1} &\sim \mathrm{Mult}(\boldsymbol{\pi}_{z_{t-1}}) \\ \boldsymbol{\theta}_k|H &\sim H \\ \mathbf{x}_t|z_t, \{\boldsymbol{\theta}_k\}_{k=1}^{\infty} &\sim F(\boldsymbol{\theta}_{z_t}) \end{aligned} \tag{6.6}$$

where it was assumed for simplicity, that there is a distinguished initial state z_0; $\boldsymbol{\beta}$ is a hyperparameter for the DP (Sethuraman, 1994) that is distributed according to the stick-breaking construction noted GEM(.); z_t is the indicator variable of the HDP-HMM that are sampled according to a multinomial distribution Mult(.); the parameters of the model are drawn independently, according to a conjugate prior distribution H; $F(\boldsymbol{\theta}_{z_t})$ is a data likelihood density, where we assume the unique parameter space of $\boldsymbol{\theta}_{z_t}$ being equal to $\boldsymbol{\theta}_k$. Suppose the observed data likelihood is a Gaussian density $\mathcal{N}(\mathbf{x}_t; \boldsymbol{\theta}_k)$ where the emission parameters $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ are respectively the mean vector $\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}_k$. According to Gelman et al. (2003); Wood and Black (2008), the prior over the mean vector and the covariance matrix is a conjugate Normal-Inverse-Wishart distribution, denoted as $\mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$, with the hyper-parameters describing the shapes and the position for each mixture densities: μ_0 is the mean of the mixtures should be, κ_0 the number of pseudo-observations supposed to be attributed, and ν_0, Λ_0 being similarly for the covariance matrix. In the generative process given in Equation (6.6), $\boldsymbol{\pi}$ is interpreted as a double-infinite transition matrix with each row taking a Chinese Restaurant Process (CRP), thus, in the HDP formulation "the group-specific" distribution, $\boldsymbol{\pi}_j$ corresponds to "the state-specific" transition where the Chinese Restaurant Franchise

(CRF) defines distributions over the next state. As a consequence it was defined the infinite state space for the Hidden Markov Model. The graphical model for the infinite Hidden Markov Model is representated in figure 6.4.

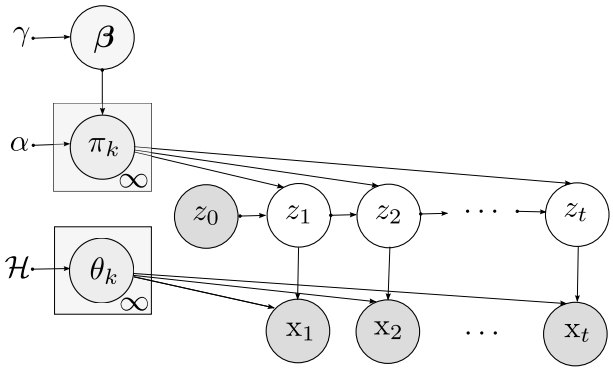


Figure 6.4: *Graphical representation of the infinite Hidden Markov Model (IHMM).*

Recalling that, the base idea of the Gibbs sampler is to estimate the posterior distributions over all the parameters from the generative process of HDP-HMM given in Equation (6.6), Beal et al. (2002) firstly considered this two level procedure of the Dirichlet Process and developed the Markov chain with the possible infinite number of states. Beal et al. (2002) considered a coupled urn model while Teh et al. (2006) developed a equivalent to the Chinese Restaurant Franchise representation of the model. Thus the infinite HMM was developed as a HDP-HMM. The inference of the infinite HMM by the Gibbs sampler was discussed by Beal et al. (2002); Teh et al. (2006) and Fox (2009) and we briefly summarized it in the Pseudo-code 8 that computes $\mathcal{O}(K)$ probabilities for each of t states, therefore it has a $\mathcal{O}(TK)$ computational complexity. The main idea to inference the HDP-HMM is to estimate the hidden states of the observed data $\mathbf{z} = (z_1, \dots z_T)$. This step needs computing two factors: the first is the conditional likelihood $p(\mathbf{x}_t|\mathbf{x}_{\setminus t}, z_t = k, \mathbf{z}_{\setminus t}, H)$ and the second factor $p(z_t|\mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha)$ computed as in Equation (6.11).

$$
\begin{aligned}
&p(z_t = k|\mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha) \propto \\
&\begin{cases}
(n_{z_t\ \ 1,k} + \alpha\beta_k)\frac{n_{k,z_{t+1}}+\alpha\beta_{z_{t+1}}}{n_{k.}+\alpha} & \text{if} \quad k \leq K,\ k \neq z_{t-1} \\
(n_{z_t\ \ 1,k} + \alpha\beta_k)\frac{n_{k,z_{t+1}}+1+\alpha\beta_{z_{t+1}}}{n_{k.}+1+\alpha} & \text{if} \quad k = z_{t-1} = z_{t+1} \\
(n_{z_t\ \ 1,k} + \alpha\beta_k)\frac{n_{k,z_{t+1}}+\alpha\beta_{z_{t+1}}}{n_{k.}+1+\alpha} & \text{if} \quad k = z_{t-1} \neq z_{t+1} \\
\alpha\beta_k\beta_{z_{t+1}} & \text{if} \quad k = K+1
\end{cases}
\end{aligned}
\tag{6.11}
$$

where n_{ij} is the number of transitions from state i to the state j, excluding the time steps t and $t-1$; $n_{.i}$ and $n_{i.}$ is the number of transition in and respectively out of state i and K is the number of distinct states in $\mathbf{z}_{\setminus t}$.

Algorithm 8 Gibbs sampler for the HDP-HMM

Inputs: The observations $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and the # of Gibbs samples n_s

Initialize a random hidden state sequence $\mathbf{z}_0 = (z_1, \ldots, z_T)$.
for $q = 1$ **to** n_s **do**
  for $t = 1$ **to** T **do**
    1. Sample the state z_t from

$$p(z_t = k|\mathbf{X}, \mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha, H) \propto p(\mathbf{x}_t|\mathbf{x}_{\setminus t}, z_t = k, \mathbf{z}_{\setminus t}, H) \; p(z_t = k|\mathbf{z}_{\setminus t}, \boldsymbol{\beta}, \alpha) \tag{6.7}$$

    2. Sample the global transition distribution

$$\boldsymbol{\beta} \propto \mathrm{Dir}(m_{.1}, \ldots, m_{.K}, \gamma) \tag{6.8}$$

    3. Sample a new transition distribution

$$\boldsymbol{\pi}_k \propto \mathrm{Dir}(n_{k1} + \alpha\beta_1, \ldots n_{kK} + \alpha\beta_K, \alpha \sum_{i=K+1}^{\infty} \beta_i) \tag{6.9}$$

    4. Sample the emission parameters $\boldsymbol{\theta}_k$.

$$\boldsymbol{\theta}_k \propto p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{z}, H, \boldsymbol{\theta}_{\setminus t}) \tag{6.10}$$

  end for
  4. Possibly update the hyper-parameters α, γ.
end for

Outputs: The states assignments $\hat{\mathbf{z}}$ and the emission parameter vector $\hat{\boldsymbol{\theta}}_k$.

Second, the global transition distribution $\boldsymbol{\beta}$ sampler is given by a Dirichlet distribution where $m_{.k}$ represents the number of clusters k, respectively one can say $m_{.k} = \sum_{j=1}^{K} m_{jk}$ (Antoniak, 1974; Teh et al., 2006). Afterwards, the transition distribution $\boldsymbol{\pi}_k$, is sampled according to the Dirichlet distribution that is followed by the sampler of the emission parameters $\boldsymbol{\theta}_k$.

Assuming that the observed data takes a Gaussian distribution, the emission parameters to be estimated are the mean vector and the covariance matrix, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. These model parameters conditional on the data $\mathbf{X}$, states $\mathbf{z}$ and the prior distribution $p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \sim \mathcal{NIW}(\mu_0, \kappa_0, \nu_0, \Lambda_0)$ are sampled according to their posterior distributions.

Finally, the hyper-parameters α and γ, because of their lack of the strong beliefs, are sampled according to a Gamma distribution Beal et al. (2002); Teh et al. (2006); Van Gael et al. (2008).

Now that, the BNP approach for the sequential data was discussed, in the next section we apply the HDP-HMM on the challenging problem of humpback whale song decomposition. This, future opens directions on deriving the HDP-HMM model to a set of parsimonious models.

6.3 Scaled application on a real-world bioacoustic data

We used the Gibbs inference algorithm for Hierarchical Dirichlet Process for Hidden Markov Model which runs for 30000 samples.

For a detailed information, the whole signal of the humpback whale song was separated by several parts of 15 seconds each. All the spectrograms of the humpback whale song and their corresponding obtained state sequence partitions, as well as the associated song are made available in the demo: http://sabiod.univ-tln.fr/workspace/IHMM_Whale_demo/. This demo highlights the interest of using the Bayesian non-parametric HMM for unsupervised structuring whale signals. Three examples of the humpback whale song, with 15 seconds duration each, are presented and discussed in this paper (see Figures (6.5), (6.6), and (6.7)).

Figure 6.5 represents the spectrogram and the corresponding state sequence partition obtained by the HDP-HMM Gibbs inference algorithm, where the selected starting time point, in the whole signal, is 60 seconds. One can see that the state 1 corresponds to the sea noise. Another thing to say is that the state 6 is not present in this time range.

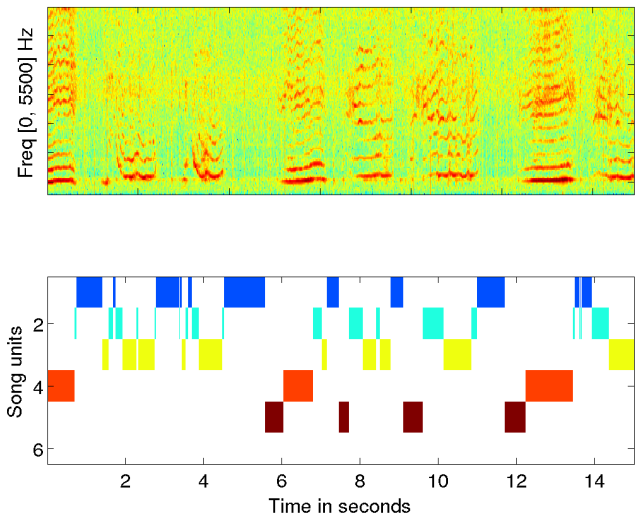


Figure 6.5: *The spectrogram of the whale song (top), starting with 60 seconds and the obtained state sequences (bottom) by the Gibbs sampler inference approach for the HDP-HMM.*

Figure 6.6 represents the spectrogram and the respective state sequence partition obtained by the HDP-HMM Gibbs inference algorithm, for the signal part starting at 255 seconds, is temporal location close to the middle

of the humpback sound recording. The sea noise, which we can see in unit 1, is predominant noise in this time step. The song unit 2, 3 and 4 song unit can be also seen in this song time range.

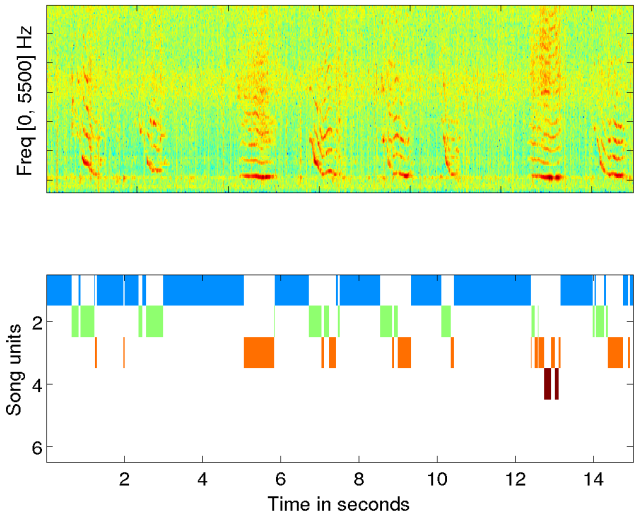


Figure 6.6: *The spectrogram of the whale song (top), starting with 255 seconds and the obtained state sequences (bottom) by the Gibbs sampler inference approach for the HDP-HMM.*

Figure 6.7 represents the spectrogram and the respective state sequences obtained by the HDP-HMM Gibbs inference algorithm, for a starting point at 495 seconds, which is close to the end of the humpback sound recording. In this time range the 6-th sound unit is the predominant one. Moreover, the sound unit 1 remains the sea noise.

All the obtained state sequences partitions fit very well the spectral patterns. We note that the estimated state 1 is the silence. The state 2 fits the up and down sweeps. State 3 fits low and high fundamental harmonics sound units, the fourth state fits for numerous harmonics sound. The fifth state is the silence, generally continued by some another sound unit, this can be due to the fact that there where not a sufficient number of Gibbs samples. For a longer learning the fifth state should be merged with the first state. Finally, the state 6 is a very well separated song unit that is a very noisy and broad sound. The analysis is discriminative on the structure.

Unlike the DPPM models applied for this complex whale song data, where it was noticed that there are a lot of states that are not used, the HDP-HMM results gives a better song structure fitting the data with 6 song units.

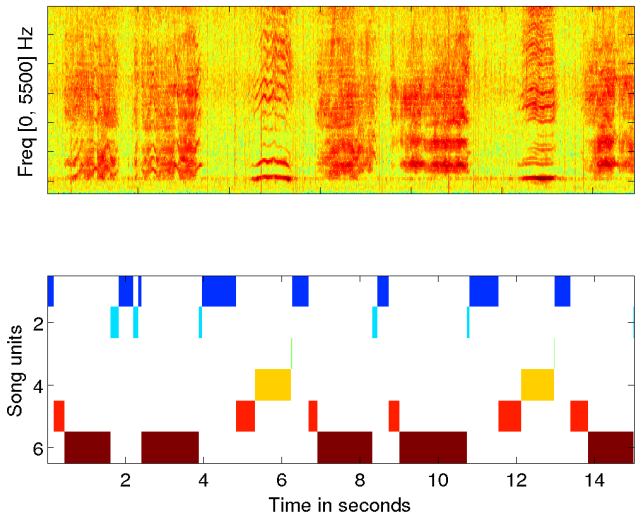


Figure 6.7: *The spectrogram of the whale song (top), starting with 495 seconds and the obtained state sequences (bottom) by the Gibbs sampler inference approach for the HDP-HMM.*

6.4 Conclusion

In this chapter we investigated an extension to the sequential case, that is the Markovian extension for the standard DPM models, in order to open feature directions to the proposed DPPM models. The infinite Hidden Markov Model, that uses a hierarchical Dirichlet Process prior over the transition matrix, named also the HDP-HMM model, was learned for the same bioacoustic data as in the previous chapter, where the DPPMs models were investigated. Indeed the obtained results provide a better fit to the data then the DPPMs, because of their exchangeability property. This study provokes possible extensions of the infinite HMM or HDP-HMM to parsimonious models, by giving eigenvalue decomposition to the covariance of the emission model components.

- Chapter 7 -

Conclusion and perspectives

7.1 Conclusions

In this thesis, we investigated the clustering based on the mixture modeling approaches. Firstly, in Chapter 2, we presented the state of the art approach on mixture modeling for model-based clustering. We focused on the Gaussian case. Then, in order to reduce the number of parameters in the mixture to be estimated, and give more flexibility in modeling the data, parsimonious mixture models were investigated. We also discussed the use of the EM algorithm which constitutes the essential feature for model fitting especially in the MLE framework. One main question also discussed in this chapter was the model selection and comparison, that is how can it be performed for the ML fitting framework.

Next, the traditional Bayesian parametric mixture modeling approaches were discussed in Chapter 3. This includes general Bayesian mixture modeling and then parsimonious Bayesian Gaussian mixture models. The Maximum A Posteriori (MAP) framework was presented as a substitution for the ML framework, allowing to avoid the problems of singularities or degeneracies. In such a context, we showed that the EM algorithm can still be used for MAP fitting, however in this work we focused on the inference using MCMC, and implemented and assessed dedicated Gibbs sampling algorithms in this Bayesian parametric framework of mixtures, particularly the parsimonious Gaussian mixtures. The Bayesian model selection and comparison was performed by the Bayes Factor, in order to select the optimal model structure.

A flexible Bayesian non-parametric alternative, to the previously investigated Bayesian and non-Bayesian parametric mixture models, was introduced in Chapter 4. We discussed Bayesian non-parametric mixture models for clustering, where the number of mixture components is estimated during the learning process. We presented our new approach, that is, the

Bayesian non-parametric parsimonious mixture models for density estimation and model-based clustering. It is based on an infinite Gaussian mixture with an eigenvalue decomposition of the cluster covariance matrix and a Dirichlet Process, or by equivalence a Chinese Restaurant Process prior. This allows deriving several flexible models and provides a well principled alternative solution of model selection encountered in the standard maximum likelihood-based and Bayesian parametric Gaussian mixture. We indeed proposed a Bayesian model selection an comparison framework to automatically select the best model structure, by using Bayes factors.

In Chapter 5, experiments carried out on simulated data highlighted that the proposed DPPMs represent a good nonparametric alternative to the standard parametric Bayesian and non-Bayesian finite mixtures. They simultaneously and accurately estimate partitions with the optimal number of clusters also inferred from the data. We also applied the proposed approach on benchmarks and real data sets, including a real challenging problem of bioacoustic data set. The possible hidden whale song units of the humpback whale signals were accurately recovered in a fully automatic way. The obtained results thus show the potential benefit of using the Bayesian parsimonious clustering models in practical applications. For example it will be used in conjunction with sparse coding decomposition of humpback whale voicing Doh (2014).

In Chapter 6, we applied the Hierarchical Dirichlet Process for Hidden Markov Model in the same challenging problem of unsupervised learning from complex bioacoustic data. Pr. Gianni Pavan (Pavia University, Italy), who is a NATO passive undersea bioacoustic expert, has analysed these results during his stay at DYNI in 2015. He validated our proposed segmentation. The obtained results are encouraging to examine of the possible extension of the sequential case.

7.2 Future works

A future work related to the proposal of the DPPM model may concern other parsimonious models such us those recently proposed by Biernacki and Lourme (2014) based on a variance-correlation decomposition of the group covariance matrices, which are stable and visualizable and have desirable properties.

The Bayesian non-parametric Markovian model (HDP-HMM) applied on a challenging bioacoustic data set has showed satisfactory results and hence opens a future direction in which we would consider the eigenvalue decomposition for the covariance matrix for the emission density of the infinite HMM. More flexible models could appear in term of different volumes, orientations and shapes for each state.

Recently, the mixture of skew-t distributions (Lee and McLachlan, 2015,

2013) received a lot of attention, these giving great performances in the clustering applications. Parsimonious skew mixture models for model-based clustering were investigated in Vrbik and McNicholas (2014). In a future work, the derivation of such models from a Bayesian non-parametric prospective would be a good alternative to deal with the problem of model selection.

Until now we have only considered the problem of clustering. A perspective of this work is to extend it to the case of model-based co-clustering (Govaert and Nadif, 2013) with block mixture models, which consists in simultaneously cluster individuals and variables, rather that only individuals. The nonparametric formulation of these models may represent a good alternative to select the number of latent blocks or co-clusters.

We also mention that the computation time for the benchmarks were reasonable due to their small number of observations, however we noticed a long computational time for the challenging bioacoustic data which contains more than 50000 individuals and can be considered from a statistical point of view as a large data set. It took around one day and half for the DPPMs and around one day for HDP-HMM. This difference may be attributed to the fact that the DPPMs Gibbs algorithm was coded in matlab while the HDP-HMM software was given with a lot of C++ routines. Thus one future work could be of course to optimize the code by using C++ routines in the DPPMs. Also different methods, to learn the DPPMs could be considered in a future toolkit developed (for example the Approximate Bayesian Computation (ABC) methods etc.) in order to reduce the learning time for the real-world data sets.

Appendix A

A.1 Prior and posterior distributions for the model parameters

Here we provide the prior and posterior distributions (used in the Gibbs sampler) for the mixture model parameters for each of the developed DPPM models. First, recall that $\mathbf{z} = (z_1, \ldots, z_n)$ denotes a vector of class labels where z_i is the class label of $\mathbf{x}_i$. Let z_{ik} be the indicator binary variable such that $z_{ik} = 1$ if $z_i = k$ (i.e when $\mathbf{x}_i$ belongs to component k). Then, let $n_k = \sum_{i=1}^{n} z_{ik}$ represents the number of data points belonging to cluster (or component) k. Finally, let $\bar{\mathbf{x}}_k = \frac{\sum_{i=1}^{n} z_{ik}\mathbf{x}_i}{n_k}$ be the empirical mean vector of cluster k, and $\mathbf{W}_k = \sum_{i=1}^{n} z_{ik}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T$ its scatter matrix.

A.1.1 Hyperparameters values

In our experiments for the multivariate parsimonious models, we choose the prior hyperparameters $\mathcal{H}$ as follows: the mean of the data $\boldsymbol{\mu}_0$, the shrinkage $\kappa_n = 0.1$, the degrees of freedom $\nu_0 = d+2$, the scale matrix $\boldsymbol{\Lambda}_0$ equal to the covariance of the data, and for the spherical models, the hyperparameter s_0^2 was taken as the greatest eigenvalue of $\boldsymbol{\Lambda}_0$.

A.1.2 Spherical models

(1) Model $\lambda\mathbf{I}$ For this spherical model, the covariance matrix, for all the mixture components, is parametrized as $\lambda\mathbf{I}$ and hence is described by the scale parameter $\lambda > 0$, which is common for all the mixture components. For this spherical model, the prior over the covariance matrix is defined through the prior over λ, for which we used a conjugate prior density, that is an inverse Gamma. For the mean vector for each of Gaussian components, we used a conjugate multivariate normal prior. The resulting prior density

is therefore a normal inverse Gamma conjugate prior:

$$\begin{aligned} \boldsymbol{\mu}_k|\lambda &\sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda\mathbf{I}/\kappa_n)\ \forall k = 1, \ldots, K \\ \lambda &\sim \mathcal{IG}(\nu_0/2, s_0^2/2) \end{aligned} \tag{A.1}$$

where $(\boldsymbol{\mu}_0, \kappa_n)$ are the hyperparamerets for the multivariate normal over $\boldsymbol{\mu}_k$ and (ν_0, s_0^2) are those for the inverse Gamma over λ. Therefore, the resulting posterior is a multivariate Normal inverse Gamma and the sampling from this posterior density is performed as follows:

$$\begin{aligned} \boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \lambda, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \lambda\mathbf{I}/(n_k + \kappa_n)) \\ \lambda|\mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IG}(\frac{\nu_0 + n}{2}, \frac{1}{2}\{s_0^2 + \sum_{k=1}^{K} \mathrm{tr}(\mathbf{W}_k) + \sum_{k=1}^{K} \frac{n_k\kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)\}) \end{aligned}$$

where the posterior mean $\boldsymbol{\mu}_n$ is equal to $\frac{n_k\bar{\mathbf{x}}_k + \kappa_n\boldsymbol{\mu}_0}{n_k + \kappa_n}$.

(2) Model $\lambda_k\mathbf{I}$ This other spherical model parametrized $\lambda_k\mathbf{I}$ is also described by the scale parameter $\lambda_k > 0$ which is different for all the mixture components. As for the previous spherical model, a normal inverse Gamma conjugate prior is used. In this situation the scale parameter λ_k will have different priors and respectively posterior distributions for each mixture component. The resulting prior density for this spherical model is a normal inverse Gamma conjugate prior:

$$\begin{aligned} \boldsymbol{\mu}_k|\lambda_k &\sim \mathcal{N}(\mu_0, \lambda_k\mathbf{I}/\kappa_n)\ \forall k = 1, \ldots, K \\ \lambda_k &\sim \mathcal{IG}(\nu_k/2, s_k^2/2)\ \forall k = 1, \ldots, K \end{aligned}$$

where $(\boldsymbol{\mu}_0, \kappa_n)$ are the hyperparamerets for the multivariate normal over $\boldsymbol{\mu}_k$ and (ν_k, s_k^2) are those for the inverse Gamma over λ_k. The set of hyperparameters $\nu_k = \{\nu_1, \ldots, \nu_k\}$ and $s_k = \{s_1 \ldots s_k\}$ are chosen to be equal, throw all the components of the mixture, to ν_0 and respectively s_0^2. Analogously, the resulting posterior is a normal inverse Gamma and the sampling for the model parameters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \lambda_1, \ldots, \lambda_K)$ is performed as follows:

$$\begin{aligned} \boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \lambda_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \lambda_k\mathbf{I}/(n_k + \kappa_n)) \\ \lambda_k|\mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IG}(\frac{\nu_k + dn_k}{2}, \frac{1}{2}\{s_k^2 + \mathrm{tr}(\mathbf{W}_k) + \frac{n_k\kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)\}). \end{aligned}$$

A.1.3 Diagonal models

(3) Model $\lambda\mathbf{A}$ The diagonal parametrization $\lambda\mathbf{A}$ of the covariance matrix is described by the volume λ (a scalar term) and a diagonal matrix $\mathbf{A}$. The parametrization $\lambda\mathbf{A}$ therefore corresponds to a diagonal matrix whose diagonal terms are a_j, $\forall j = 1, \ldots d$. The prior normal inverse Gamma conjugate prior density is given as follows:

$$\begin{aligned} \boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k/\kappa_n)\ \forall k = 1, \ldots, K \\ a_j &\sim \mathcal{IG}(r_j/2, p_j/2)\ \forall j = 1 \ldots d \end{aligned}$$

where the set of parameters r_j, p_j are considered to be equal $\forall j = 1 \dots d$ to ν_0 and respectively s_k^2. The resulting posterior for the model parameters takes the following form:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \boldsymbol{\Sigma}_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_k/(n_k + \kappa_n)) \\
a_j|\mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IG}(\frac{n + \nu_k + K(d+1) - 2}{2}, \frac{\mathrm{diag}(\sum_{k=1}^{K} \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \boldsymbol{\Lambda}_k)}{2})
\end{aligned}$$

where the posterior mean $\boldsymbol{\mu}_n = \frac{n_k \bar{\mathbf{x}}_k + \kappa_n \boldsymbol{\mu}_0}{n_k + \kappa_n}$.

(4) Model $\lambda_k \mathbf{A}$ This diagonal model, analogous to the previous one, but with different volume $\lambda_k > 0$ for each component of the mixture, takes the parametrization $\lambda_k \mathbf{A}$. In this situation, the normal prior density for the mean remains the same and the inverse Gamma prior density for the volume parameter λ_k is given as follows:

$$\lambda_k \sim \mathcal{IG}(r_k/2, p_k/2) \ \forall j = 1 \dots K$$

where the set of hyperparamerets for the scale parameter λ_k, $r_k = \{r_1, \dots, r_K\}$ and $p_k = \{p_1, \dots, p_k\}$ are considered to be equal, for all mixture components, to respectively ν_0 and s_k^2. The resulting posterior distributions over the parameters of the model are given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \boldsymbol{\Sigma}_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_k/(n_k + \kappa_n)) \\
a_j|\mathbf{X}, \mathbf{z}, \lambda_k, \mathcal{H} &\sim \mathcal{IG}(\frac{n + \nu_k + Kd + 1}{2}, \frac{\mathrm{diag}(\sum_{k=1}^{K} \lambda_k^{-1}(\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \boldsymbol{\Lambda}_k))}{2}) \\
\lambda_k|\mathbf{X}, \mathbf{z}, \mathbf{A}, \mathcal{H} &\sim \mathcal{IG}(\frac{r_k + n_k d}{2}, \frac{p_k + \mathrm{tr}(\mathbf{A}^{-1}(\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \boldsymbol{\Lambda}_k))}{2}).
\end{aligned}$$

A.1.4 General models

(5) Model $\lambda \mathbf{DAD}^T$ The first general model has the $\lambda \mathbf{DAD}^T$ parametrization, where the covariance matrices have the same volume $\lambda > 0$, orientation $\mathbf{D}$ and shape $\mathbf{A}$ for all the components of the mixture. This is equivalent, in the literature, to the model where the covariance $\boldsymbol{\Sigma}$ is considered equal throw all the components of the mixture. The resulting conjugate normal inverse Wishart prior over the parameters $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma})$ is given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k|\boldsymbol{\Sigma} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_n) \ \forall k = 1, \dots, K \\
\boldsymbol{\Sigma} &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0)
\end{aligned}$$

where $(\boldsymbol{\mu}_0, \kappa_n)$ are the hyperparameters for the multivariate normal prior over $\boldsymbol{\mu}_k$ and $(\nu_0, \boldsymbol{\Lambda}_0)$ are hyperparameters for the inverse Wishart prior ($\mathcal{IW}$) over the covariance matrix $\boldsymbol{\Sigma}$ that is common to all the components of the mixture. The posterior of the model parameters $(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma})$ for this general model is given by:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \lambda_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/(n_k + \kappa_n)) \\
\boldsymbol{\Sigma}|\mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IW}(\nu_0 + n, \boldsymbol{\Lambda}_0 + \sum_{k=1}^{K}\{\mathbf{W}_k + \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T\}).
\end{aligned}$$

(6) Model $\lambda_k \mathbf{DAD}^T$ The second parsimonious model from the general family has the parametrization $\lambda_k \mathbf{DAD}^T$, where the volume λ_k of the covariance differs from one mixture component to another, but the orientation $\mathbf{D}$ and the shape $\mathbf{A}$ are the same for all the mixture components. This parametrization can thus be simplified as $\lambda_k \mathbf{\Sigma}_0$, where the parameter $\mathbf{\Sigma}_0 = \mathbf{DAD}^T$. This general model has therefore a Normal prior distribution over the mean, an inverse Gamma prior distribution over the scale parameter λ_k and an inverse Wishart prior distribution over the matrix $\mathbf{\Sigma}_0$ that controls the orientation and the shape for the mixture components. The conjugate prior for the mixture parameters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \lambda_1, \ldots, \lambda_K, \mathbf{\Sigma}_0)$ are thus given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k | \lambda_k, \mathbf{\Sigma}_0 &\sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda_k \mathbf{\Sigma}_0 / \kappa_n) \ \forall k = 1, \ldots, K \\
\lambda_k &\sim \mathcal{IG}(r_k/2, p_k/2) \ \forall k = 2, \ldots, K \\
\mathbf{\Sigma}_0 &\sim \mathcal{IW}(\nu_0, \mathbf{\Lambda}_0)
\end{aligned}$$

where λ_1 is supposed to be equal to 1 (to make the model identifiable), the hyperparameters $\{r_1, \ldots, r_K\}$ and $\{p_1 \ldots p_K\}$ are supposed to be equal to respectively ν_0 and s_k^2 for each of the mixture components. The resulting posterior over the parameters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \lambda_1, \ldots, \lambda_K, \mathbf{\Sigma}_0)$ of this model is given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k | \mathbf{X}, \mathbf{z}, \lambda_k, \mathbf{\Sigma}_0, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \lambda_k \mathbf{\Sigma}_0 / (n_k + \kappa_n)) \\
\lambda_k | \mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IG}(\frac{r_k + n_k d}{2}, \frac{1}{2}\{p_k + \mathrm{tr}(\mathbf{W}_k \mathbf{\Sigma}_0^{-1}) + \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1} (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)\}) \\
\mathbf{\Sigma}_0 | \mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IW}(\nu_0 + n, \mathbf{\Lambda}_0 + \sum_{k=1}^{K} \{\frac{\mathbf{W}_k}{\lambda_k} + \frac{n_k \kappa_n}{\lambda_k (n_k + \kappa_n)}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T (\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)\}).
\end{aligned}$$

(7) Model $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ This other general model $\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ is parametrized by the scalar parameter (the volume) λ and the shape diagonal matrix $\mathbf{A}$. This model parametrization can therefore be summarized to the $\mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ parametrization, by including λ in a resulting diagonal matrix $\mathbf{A}$, whose diagonal elements $a_1, \ldots, a_d$. The prior density over the mean is normal, the one over the orientation matrix $\mathbf{D}_k$ is inverse Wishart, and the one over each of the diagonal elements $a_j, \ \forall j = 1 \ldots d$ of the matrix $\mathbf{A}$ is an inverse Gamma. The conjugate prior for this general model is therefore as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k | \mathbf{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_k / \kappa_n) \ \forall k = 1, \ldots, K \\
a_j &\sim \mathcal{IG}(r_j/2, p_j/2) \ \forall j = 1 \ldots d
\end{aligned}$$

The hyperparameters r_j and p_j for the $\lambda \mathbf{A}$, are considered to be the same $\forall j = 1 \ldots d$ and are respectively equal to ν_0 and s_k^2. The resulting posterior for the model parameters takes the following form:

$$\begin{aligned}
\boldsymbol{\mu}_k | \mathbf{X}, \mathbf{z}, \mathbf{\Sigma}_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \mathbf{\Sigma}_k / (n_k + \kappa_n)) \\
a_j | \mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IG}(\frac{n + \nu_k + K(d+1) - 2}{2}, \frac{\mathrm{diag}(\sum_{k=1}^{K} \mathbf{D}_k^T (\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \mathbf{\Lambda}_k) \mathbf{D}_k)}{2}).
\end{aligned}$$

The parameters, that controls the orientation of the covariance, $\mathbf{D}_k$, have the same inverse Wishart posterior distribution as the general covariance

matrix:

$$\mathbf{D}_k|\mathbf{X},\mathbf{z},\mathcal{H} \sim \mathcal{IW}(n_k+\nu_k, \mathbf{\Lambda}_k + \mathbf{W}_k + \frac{n_k\kappa_n}{n_k+\kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T)$$

And as mentioned above the covariance matrix $\mathbf{\Sigma}_k$ for this model will be formed as $\text{diag}(a_j)\mathbf{D}_k$.

(8) Model $\lambda\mathbf{D}\mathbf{A}_k\mathbf{D}^T$(*) Another general model with the $\lambda\mathbf{D}\mathbf{A}_k\mathbf{D}^T$ parametrization is given. In this situation the volume parameter λ, that is equal, and the shape $\mathbf{A}_k$, that varies for all mixture components are taken not to be separated, thus the parametrization of this model is given by $\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k$, with the parameter $\mathbf{D}_k$, the cluster orientation. For this model the diagonal matrix $\mathbf{A}_k$ has the diagonal terms equal to $(1, a_{2k}, a_{3k}, \ldots, a_{dk})$ $\forall k = 1, \ldots, K$. The prior density for the diagonal elements of $\mathbf{A}_k$ is an inverse Gamma and is supposed as follows. Suppose a inverse Gamma prior for λ.

$$\lambda \sim \mathcal{IG}(\nu_0/2, s_0^2/2)$$

where (ν_0, s_0^2) are hyperparamerets of the inverse Gamma density. The resulting prior for the $\mathbf{A}_k$, $\forall k = 1, \ldots, K$ can be given by:

$$\lambda a_{tk}|\lambda \sim \mathcal{IG}(r_{tk}/2, p_{tk}/2)\ \forall j = 1, \ldots, d\ \forall k = 1, \ldots, K$$

where the hyperparameters set (r_{tk}, p_{tk}) is supposed to be equal to ν_0 and respectively s_0^2. The resulting posterior for the model parameters λa_{tk} and $\mathbf{D}$ are similar to the general model $\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^t$. But for now, in place of simulating the $\mathbf{A}_k$, the $\lambda\mathbf{A}_k$ is simulated, thus a posterior distribution over λ is given as follows:

$$\lambda|\mathbf{X},\mathbf{z},\mathcal{H} \sim (\frac{\nu_0+n}{2}, \frac{1}{2}\{s_0^2 + \sum_{k=1}^{K}\text{tr}(W_k) + \sum_{k=1}^{K}\frac{n_k\kappa_n}{n_k+\kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T\}) \tag{A.2}$$

(9) Model $\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^t$(*) In this case the model takes the parametrization $\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^t$. This consists of different volume λ_k and shape $\mathbf{A}_k$, but the same orientation $\mathbf{D}$ over the mixture components. In this situation, the separation between the volume and the shapes are not needed, therefore the parametrization of this model is supposed to be $\mathbf{D}\mathbf{A}_k\mathbf{D}^t$, where the first term of the diagonal $\mathbf{A}_k$ is not equal to one. The prior density over the mean is normal, the one over the diagonal terms of the matrix $\mathbf{A}_k$ is inverse Gamma and the prior density for the matrix $\mathbf{D}$, that is the cluster orientation, is an inverse Wishart. The conjugate prior for this general model is therefore as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{D},\mathbf{A}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{D}\mathbf{A}_k\mathbf{D}^T/\kappa_n)\ \forall k = 1, \ldots, K \\
a_{tk} &\sim \mathcal{IG}(r_{tk}/2, p_{tk}/2)\ \forall j = 1 \ldots d\ \forall k = 1, \ldots, K \\
\mathbf{D} &\sim \mathcal{IW}(\nu_0, \mathbf{I})
\end{aligned}$$

where (r_{tk}, p_{tk}), are hyperparameters for the inverse Gamma prior density. The hyperparameters $(r_{tk}$ and p_{tk}, are considered to be the same $\forall j = 1 \dots d,\ k = 1 \dots K$ and are respectively equal to ν_0 and s_k^2. The resulting posterior for the model parameters takes the following form:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \mathbf{D}, \mathbf{A}_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \frac{\boldsymbol{\Sigma}_k}{n_k + \kappa_n}) \\
a_{tk}|\mathbf{X}, \mathbf{z}, \mathbf{D}, H &\sim \mathcal{IG}(\frac{r_{tk} + n_k}{2}, \frac{\mathrm{diag}(\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{D}^T W_k \mathbf{D})}{2}) \\
\mathbf{D}|\mathbf{X}, \mathbf{z}, \mathbf{A}_k, H &\sim \mathrm{diag}(\mathbf{D}\mathbf{D}^T)^{-(\nu_0+d+1)/2} \exp\left\{ -\frac{1}{2}\mathrm{tr}\left(\sum_{k=1}^{K_+} \mathbf{A}_k^{-1}\mathbf{D}^T[(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T \frac{n_k \kappa_n}{n_k + \kappa_n} + \mathbf{W}_k] \right) \right\}
\end{aligned}$$

where the posterior mean $\boldsymbol{\mu}_n$ is equal to $\frac{n_k \bar{\mathbf{x}}_k + \kappa_n \boldsymbol{\mu}_0}{n_k + \kappa_n}$.

(10) Model $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ The third considered parsimonious model for the general family, is the one with the parametrization $\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}_k^T$ of the covariance matrix, and is analogous to the previous model, but for this one, the scale λ_k of the covariance (the cluster volume) differs for each component of the mixture. The prior over each of the scale parameters $\lambda_1 \dots \lambda_K$ is an inverse Gamma prior :

$$\lambda_k \sim \mathcal{IG}(r_k/2, p_k/2)\ \forall k = 1, \dots, K.$$

The set of hyperparameters $r_k = \{r_1, \dots r_K\}$ and $p_k = \{p_1, \dots p_K\}$ are considered equal between the components of the mixture and are taken equal to respectively ν_0 and s_k^2. The resulting posterior distributions over the parameters of the model are given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k|\mathbf{X}, \mathbf{z}, \boldsymbol{\Sigma}_k, \mathcal{H} &\sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_k/(n_k + \kappa_n)) \\
a_j|\mathbf{X}, \mathbf{z}, \lambda_k, \mathbf{D}_k, \mathcal{H} &\sim \mathcal{IG}(\frac{n + \nu_k + Kd + 1}{2}, \frac{\mathrm{diag}(\sum_{k=1}^{K} \lambda_k^{-1}\mathbf{D}_k^T(\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \boldsymbol{\Lambda}_k)\mathbf{D}_k)}{2}) \\
\mathbf{D}_k|\mathbf{X}, \mathbf{z}, \mathcal{H} &\sim \mathcal{IW}(n_k + \nu_k, \boldsymbol{\Lambda}_k + \mathbf{W}_k + \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T) \\
\lambda_k|\mathbf{X}, \mathbf{z}, \mathbf{D}_k, \mathbf{A}_k, \mathcal{H} &\sim \mathcal{IG}(\frac{r_k + n_k d}{2}, \frac{p_k + \mathrm{tr}(\mathbf{D}_k \mathbf{A}^{-1} \mathbf{D}_k^T(\frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T + \mathbf{W}_k + \boldsymbol{\Lambda}_k))}{2}).
\end{aligned}$$

(11) Model $\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$(*) For this situation, the model has the parametrization $\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$. This can be simplified by the $\lambda \boldsymbol{\Sigma}_{0k}$ parametrization, with the multivariate normal prior density for the mean vector, the inverse Gamma prior density for λ, and the inverse Wishart prior density for the $\boldsymbol{\Sigma}_{0k}$. The considered prior density are given as follows:

$$\begin{aligned}
\boldsymbol{\mu}_k|\lambda, \boldsymbol{\Sigma}_{0k} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \lambda \boldsymbol{\Sigma}_{0k}/\kappa_n)\ \forall k = 1, \dots, K \\
\lambda &\sim \mathcal{IG}(nu_0/2, s_0^2/2)\ \forall j = 1 \dots d\ \forall k = 1, \dots, K \\
\boldsymbol{\Sigma}_{0k} &\sim \mathcal{IW}(\nu_k, \boldsymbol{\Lambda}_k)\ \forall k = 1, \dots, K
\end{aligned}$$

The resulted posterior distributions for the mean vector $\boldsymbol{\mu}_k$ and the matrix $\boldsymbol{\Sigma}_{0k}$ are considered to be the same as in the full-GMM model with

$\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ parametrization. $\boldsymbol{\Sigma}_k$ will be replaced by $\boldsymbol{\Sigma}_{0k}$. For the λ parameter, the posterior distribution is given as follows:

$$\lambda|\mathbf{X}, \mathbf{z}, \boldsymbol{\Sigma}_{0k}, \boldsymbol{\mu}_k \mathcal{H} \sim \mathcal{IG}(\frac{\nu_0 + n}{2}, \frac{1}{2}s_0^2 + \sum_k \mathrm{tr}(\mathbf{W}_k) + \sum_k \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0))$$

(12) Model $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ Finally, the more general model is the standard one with $\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$ parametrization. This model is also known as the full covariance model $\boldsymbol{\Sigma}_k$. The volume λ_k, the orientation $\mathbf{D}_k$, and the shape $\mathbf{A}_k$ differ for each component of the mixture. In this situation, the prior density for the mean is normal and the one for the covariance matrix is an inverse Wishart, which leads to the following conjugate normal inverse Wishart prior density:

$$\begin{aligned} \boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k/\kappa_n) \ \forall k = 1, \ldots, K \\ \boldsymbol{\Sigma}_k &\sim \mathcal{IW}(\nu_k, \boldsymbol{\Lambda}_k) \ \forall k = 1, \ldots, K \end{aligned}$$

where $(\boldsymbol{\mu}_0, \kappa_n)$ and $(\nu_k, \boldsymbol{\Lambda}_k)$ are respectively the hyperparamerets for respectively normal prior density over the mean and the inverse Wishart prior density over the covariance matrix. The resulting posterior over the model parameters $(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k)$ is given as follows:

$$\boldsymbol{\Sigma}_k|\mathbf{X}, \mathbf{z}, \mathcal{H} \sim \mathcal{IW}(n_k + \nu_k, \boldsymbol{\Lambda}_k + \mathbf{W}_k + \frac{n_k \kappa_n}{n_k + \kappa_n}(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_k - \boldsymbol{\mu}_0)^T).$$

Appendix B

B.1 Multinomial distribution

Suppose the components $\theta_k = \{0, 1\}$ such that $\sum_k \theta_k = 1$, the following discrete distribution is given as a multivariate generalization of the Bernoulli distribution. The pdf of multinomial distribution is given by the following:

$$p(\boldsymbol{\theta}) = \prod_{k=1}^{K} \mu_k^{\theta_k} \tag{B.1}$$

where $\boldsymbol{\theta}$ is a K dimensional binary variable with θ_k components.

B.2 Normal-Inverse Wishart distribution

Suppose nether the mean vector, neither the covariance matrix of the GMM are known. The normal inverse Wishart distribution is then supposed for the model parameters.

$$\begin{aligned}
\boldsymbol{\Sigma}_k &\sim \mathcal{IW}(\nu_0, \boldsymbol{\Lambda}_0) \\
&= 2\pi \left| \frac{\boldsymbol{\Sigma}_k}{\kappa_0} \right|^{1/2} \exp\{-\frac{\kappa_0}{2}(\mathbf{x}_i - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_0)\} && \text{(B.2)} \\
\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, \frac{\boldsymbol{\Sigma}_k}{\kappa_0}) \\
&= \frac{|\boldsymbol{\Lambda}_0|^{\nu/2}}{2^{\frac{\nu d}{2}}\Gamma_d(\nu/2)} |\boldsymbol{\Sigma}_k|^{-\frac{\nu+d+1}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_0 \boldsymbol{\Sigma}_k^{-1})\} && \text{(B.3)}
\end{aligned}$$

with normal distribution $\mathcal{N}$ and the Inverse-Wishart distribution $\mathcal{IW}$.

The log form for this distribution is given respectively as follows:

$$\begin{aligned}
\log p(\boldsymbol{\Sigma}_k|\boldsymbol{\Lambda}_0,\nu) &= \log\left(\frac{|\boldsymbol{\Lambda}_0|^{\nu/2}}{2^{\frac{\nu d}{2}}\Gamma_d(\nu/2)}|\boldsymbol{\Sigma}_k|^{-\frac{\nu+d+1}{2}}\exp\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}_k^{-1})\}\right) \\
&= \frac{\nu}{2}\log|\boldsymbol{\Lambda}_0| - \frac{\nu d}{2}\log(2) - \log(\Gamma_d(\nu/2)) - \\
&\quad -\frac{\nu+d+1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}_k^{-1}) \qquad \text{(B.4)}
\end{aligned}$$

where $\boldsymbol{\Lambda}_0$ and ν are hyperparameters representing the positive definite matrix d x d and the degree of freedom $\nu > d-1$. $\Gamma_d(.)$ represents the multivariate gamma function that is a generalization of gamma distribution and is defined by the Equation (B.5)

$$\Gamma_d(x) = \pi^{d(d-1)/4}\prod_{i=1}^{d}\Gamma[x+(1-j)/2] \qquad \text{(B.5)}$$

$$\begin{aligned}
\log p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k,\boldsymbol{\mu}_0,\kappa_0) &= \log\left(2\pi\left|\frac{\boldsymbol{\Sigma}_k}{\kappa_0}\right|^{1/2}\exp\{-\frac{\kappa_0}{2}(\mathbf{x}_i-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_0)\}\right) \\
&= \log(2\pi) + \frac{1}{2}\log\left|\frac{\boldsymbol{\Sigma}_k}{\kappa_0}\right| - \\
&\quad -\frac{\kappa_0}{2}(\mathbf{x}_i-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_0) \qquad \text{(B.6)}
\end{aligned}$$

B.3 Dirichlet distribution

The Dirichlet distribution, that is a multivariate generalization of the beta distribution, is parametrized by a vector $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)$ of a positive real numbers. The pdf of the Dirichlet distribution is given by the following:

$$f(\theta_1,\theta_2,\ldots,\theta_K;\alpha_1,\alpha_2,\ldots,\alpha_K) = \frac{\Gamma(\sum_{k=1}^{K}\theta_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \qquad \text{(B.7)}$$

where $\sum_{k=1}^{K}\theta_k = 1$ and $0 < \theta_k < 1$.

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. 2, 6, 27

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. 2, 6, 62

J. Almhana, Z. Liu, V. Choulakian, and R. McGorman. A recursive algorithm for gamma mixture models. In *Communications, 2006. ICC '06. IEEE International Conference on*, volume 1, pages 197–202, June 2006. 11

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and MichaelI. Jordan. An Introduction to MCMC for Machine Learning. *Machine Learning*, 50(1-2):5–43, 2003. 71

Charles E. Antoniak. Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174, 1974. 2, 6, 61, 62, 68, 75, 76, 118

W.W.L. Au, A. Frankel, D.A. Helweg, and D.H. Cato. Against the humpback whale sonar hypothesis. *Oceanic Engineering, IEEE Journal of*, 26 (2):295–300, April 2001. 97

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985. 11

A. Azzalini and A. W. Bowman. A look at some data on the Old Faithful geyser. *Applied Statistics*, pages 357–365, 1990. 22, 89

C. Scott Baker and Louis M. Herman. Aggressive behavior between humpback whales (Megaptera novaeangliae) wintering in Hawaiian waters. *Canadian Journal of Zoology*, 62(10):1922–1937, 1984. 97

J. D. Banfield and A. E. Raftery. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49(3):803–821, 1993. 1, 2, 5, 6, 10, 11, 14, 15, 18, 27, 28, 34, 37, 49, 60, 72

Marius Bartcus, Faicel Chamroukhi, Joseph Razik, and Hervé Glotin. Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture. In *Proceedings of the Neural Information Processing Systems (NIPS), workshop on Neural Information Processing Scaled for Bioacoustics: NIPS4B*, pages 205–211, Nevada, USA, December 2013. 3, 7, 99

Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Clustering Bayésien Parcimonieux Non-Paramétrique. In *Proceedings of 14èmes Journées Francophones Extraction et Gestion des Connaissances (EGC), Atelier CluCo: Clustering et Co-clustering*, pages 3–13, Rennes, France, Janvier 2014. 3, 7

Marius Bartcus, Faicel Chamroukhi, and Hervé Glotin. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, July 2015. 112

S. Basu and S. Chib. Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models. *Journal of the American Statistical Association*, 98:224–235, 2003. 7, 50

Matthew J. Beal, Zoubin Ghahramani, and Carl E. Rasmussen. The infinite hidden markov model. In *Machine Learning*, pages 29–245. MIT Press, 2002. 4, 8, 109, 112, 116, 117, 118

H. Bensmail and Jacqueline J. Meulman. Model-based Clustering with Noise: Bayesian Inference and Estimation. *Journal of Classification*, 20 (1):049–076, 2003. 2, 6, 34, 38, 43, 46, 49, 50, 52, 53, 60, 72, 159

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. 2, 6, 34, 35, 38, 43, 46, 49, 50, 52, 53, 60, 72, 73, 80, 81, 87, 159

Halima Bensmail. *Modèles de régularisation en discrimination et classification bayésienne*. PhD thesis, Université Paris 6, 1995. 2, 6, 14, 38, 46, 49, 50, 51, 52, 53, 60, 81, 159

Halima Bensmail and Gilles Celeux. Regularized Gaussian Discriminant Analysis through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996. 2, 6, 14, 25, 60, 81

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. 2, 6, 27, 28

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41:561–575, 2003. 21

Christophe Biernacki. *Choix de modèles en classification*. PhD thesis, Université de Technologie de Compiègne, 1997. 27

Christophe Biernacki. Initializing EM Using the Properties of Its Trajectories in Gaussian Mixtures. *Statistics and Computing*, 14(3):267–279, August 2004. 21

Christophe Biernacki and Gérard Govaert. Choosing models in model-based clustering and discriminant analysis. Technical Report RR-3509, INRIA, Rocquencourt, 1998. 27, 50

Christophe Biernacki and Alexandre Lourme. Stable and visualizable gaussian parsimonious clustering models. *Statistics and Computing*, 24(6): 953–969, 2014. 124

D. Blackwell and J. MacQueen. Ferguson Distributions Via Polya Urn Schemes. *The Annals of Statistics*, 1:353–355, 1973. 62, 64

David M. Blei and Michael I. Jordan. Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144, 2006. 61

Dankmar Böhning. *Computer-Assisted Analysis of Mixtures and Applications. Meta-Analysis, Disease Mapping, and Others*. Chapman & Hall, Boca Raton, 1999. 10

Charles Bouveyron. *Modélisation et classification des données de grande dimension: application à l'analyse d'images*. PhD thesis, Université Joseph Fourier, September 2006. 2, 6, 12

Charles Bouveyron and Camille Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71(C):52–78, 2014. 2, 6, 12, 15

H. Bozdogan. Determining the number of component clusters in the standard multi-variate normal mixture model using model-selection criteria. Technical report, Quantitative Methods Department, University of Illinois at Chicago, June 1983. 27, 28

N. A. Campbell and R. J. Mahon. A multivariate study of variation in two species of rock crab of genus Leptograpsus. *Australian Journal of Zoology*, 22:417–425, 1974. 91

Bradley P. Carlin and Siddhartha Chib. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B*, 57(3):473–484, 1995. 7, 50

George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):pp. 167–174, 1992. 44

George Casella and Christian P. Robert. Rao-Blackwellisation of sampling schemes. *Biometrika*, 83(1):81–94, March 1996. 71

G. Celeux and J. Diebolt. The SEM algorithm a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985. 17, 21

G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14: 315–332, 1992. 17, 21

G. Celeux and G. Govaert. Gaussian Parsimonious Clustering Models. *Pattern Recognition*, 28(5):781–793, 1995. 1, 2, 3, 5, 6, 7, 11, 14, 15, 16, 17, 24, 25, 34, 37, 40, 53, 60, 72, 80, 81, 85

G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. Technical Report RR-2514, The French National Institute for Research in Computer Science and Control (INRIA), 1995. 17

Gilles Celeux. Bayesian Inference for Mixture: The Label Switching Problem. In Roger Payne and Peter Green, editors, *COMPSTAT*, pages 227–232. Physica-Verlag HD, 1998. 77

Gilles Celeux, Didier Chauveau, and Jean Diebolt. Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314, 1996. 17

Gilles Celeux, Merrilee Hurn, and Christian P. Robert. Computational and Inferential Difficulties With Mixture Posterior Distributions. *Journal of the American Statistical Association*, 95:957–970, 1999. 77

Faicel Chamroukhi, Marius Bartcus, and Hervé Glotin. Bayesian Non-Parametric Parsimonious Clustering. In *Proceedings of 22nd European Symposium on Artifcial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Bruges, Belgium, April 2014a. 3, 7

Faicel Chamroukhi, Marius Bartcus, and Hervé Glotin. Bayesian Non-Parametric Parsimonious Gaussian Mixture for Clustering. In *Proceedings of 22nd International Conference on Pattern Recognition (ICPR)*, Stockholm, Sweden, August 2014b. 3, 7

Faicel Chamroukhi, Marius Bartcus, and Hervé Glotin. Dirichlet Process Parsimonious Mixture for clustering. January 2015. Preprint, 35 pages, available online arXiv:501.03347. Submitted to Patter Recognition - Elsevier. 3, 7

S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. 52

Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, New York, 2008. 27

Abhijit Dasgupta and Adrian E. Raftery. Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering. *Journal of the American Statistical Association*, 93(441):pp. 294–302, 1998. 28

N. Day. Estimation of components of a mixture of normal distribution. *Biometrica*, 56:463–474, 1969. 11, 34

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. 3, 7, 17, 20

Bernard Desgraupes. Clustering Indices. Technical report, University Paris Ouest Lab Modal'X, 2013. 47

Jean Diebolt and Christian P. Robert. Estimation of Finite Mixture Distributions through Bayesian Sampling. *Journal of the Royal Statistical Society. Series B*, 56(2):363–375, 1994. 2, 3, 6, 7, 34, 35, 38, 43, 44, 49, 74

Yann Doh. *Nouveaux modèles d'estimation monophone de distance et d'analyse parcimonieuse - Applications sur signaux transitoires et stationnaires bioacoustiques à l'échelle*. PhD thesis, Université de Toulon, 17 décembre 2014. 124

Michael D. Escobar. Estimating Normal Means with a Dirichlet Process Prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994. 68

Michael D. Escobar and Mike West. Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1994. 34, 43, 75

Michael Evans, Irwin Guttman, and Ingram Olkin. Numerical aspects in estimating the parameters of a mixture of normal distributions. *Journal of Computational and Graphical Statistics*, 1(4):351–365, 1992. 34

Thomas S. Ferguson. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):209–230, 1973. ISSN 00905364. 2, 6, 61, 62, 63, 66, 68, 112, 113

Thomas S. Ferguson. Prior Distributions on Spaces of Probability Measures. *Ann. Statist.*, 2(4):615–629, 07 1974. 62

R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(7):179–188, 1936. 23, 95

E.B. Fox. *Bayesian Nonparametric Learning of Complex Dynamical Phenomena*. Phd. thesis, MIT, Cambridge, MA, 2009. 4, 8, 112, 115, 116, 117

Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. An HDP-HMM for systems with state persistence. In *ICML 2008: Proceedings of the 25th international conference on Machine learning*, pages 312–319, New York, NY, USA, 2008. ACM. 4, 8, 112, 115, 116

C. Fraley and A. E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8): 578–588, August 1998a. 1, 5, 10, 11, 28, 34

C. Fraley and A. E. Raftery. Mclust: Software for model-based cluster and discriminant analysis, 1998b. 14, 16, 41

C. Fraley and A. E. Raftery. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association*, 97:611–631, 2002. 2, 6, 10, 14, 60

C. Fraley and A. E. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *Journal of Classification*, 24(2): 155–181, September 2007a. ISSN 0176-4268. 2, 3, 6, 7, 14, 28, 34, 35, 37, 38, 39, 40, 41, 42, 50, 60, 72, 73, 80

Chris Fraley and Adrian Raftery. Model-based methods of classification: Using the mclust software in chemometrics. *Journal of Statistical Software*, 18(6):1–13, 1 2007b. ISSN 1548-7660. 14, 16, 41

Chris Fraley and Adrian E. Raftery. Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. Technical Report 486, Departament of Statistics, University of Washington Seattle, 2005. 2, 3, 6, 7, 14, 28, 34, 35, 37, 38, 39, 40, 41, 42, 60, 72, 73

A. S. Frankel, C. W. Clark, L. M. Herman, and C. M. Gabriele. Spatial distribution, habitat utilization, and social interactions of humpback whales, Megaptera novaeangliae, off Hawai'i, determined using acoustic and visual techniques. *Canadian Journal of Zoology*, 73(6):1134–1146, 1995. 97

L.N. Frazer and E. Mercado. A sonar model for humpback whale song. *Oceanic Engineering, IEEE Journal of*, 25(1):160–182, January 2000. 97

David A. Freedman. On the asymptotic behavior of bayes estimates in the discrete case ii. *The Annals of Mathematical Statistics*, 36(2):454–456, 1965. 62

Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer series in statistics. Springer, New York, 2006. 1, 5, 10

Ellen C. Garland, Anne W Goldizen, Melinda L. Rekdahl, Rochelle Constantine, Claire Garrigue, Nan Daeschler Hauser, M. Michael Poole, Jooke Robbins, and Michael J. Noad. Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current Biology*, 21(8): 687–691, 2011. 97

A. E. Gelfand and D. K. Dey. Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B*, 56 (3):501–514, 1994. 7, 50, 52

Alan E. Gelfand and Adrian F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85(410):398–409, June 1990. 44, 45, 74

Alan E. Gelfand, Susan E. Hills, Amy Racine-Poon, and Adrian F. M. Smith. Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association*, 85(412):972–985, December 1990. 44

Andrew Gelman and Gary King. Estimating the electoral consequences of legislative redistricting. *Journal of the American Statistical Association*, 85(410):274–282, June 1990. 34

Andrew Gelman and Donald B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):pp. 457–472, 1992. 45

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003. 34, 36, 41, 45, 116

Stuart Geman and Donald Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 6(6):721–741, November 1984. 44, 74

C. Geyer. Markov Chain Monte Carlo maximum likelihood. In *Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991. 3, 7, 43

Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992. 45

Zoubin Ghahramani and Geoffrey E. Hinton. The EM Algorithm for Mixtures of Factor Analyzers. Technical report, University of Toronto, 1997. 11

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.
This book thoroughly summarizes the uses of MCMC in Bayesian analysis. It is a core book for Bayesian studies. 3, 7, 43, 44

Gérard Govaert and Mohamed Nadif. *Co-Clustering*. Computer engineering series. Wiley, November 2013. 256 pages. 125

Peter J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82:711–732, 1995. 38

Peter J. Green and Sylvia Richardson. Modelling heterogeneity with and without the dirichlet process. *Scandinavian Journal of Statistics*, 28(2): 355–375, 2001. ISSN 1467-9469. 70

Arjun K. Gupta, Graciela González-Farías, and J.Armando Domínguez-Molina. A multivariate skew normal distribution. *Journal of Multivariate Analysis*, 89(1):181 – 190, 2004. 11

Dilan Görür. *Nonparametric Bayesian discrete latent variable models for unsupervised learning*. PhD thesis, Berlin Institute of Technology, 2007. 71

Dilan Görür and Carl Edward Rasmussen. Dirichlet Process Gaussian Mixture Models: Choice of the Base Distribution. *Journal of Computer Science and Technology*, 25(4):653–664, 2010. doi: 10.1007/s11390-010-9355-8. 70

Peter Hall, S. Marron J., and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society Series B*, 67(3):427–444, 2005. 14

John Michael Hammersley and David Christopher Handscomb. *Monte Carlo methods*. Monographs on statistics and applied probability. Chapman and Hall, London, 1964. 51

Trevor Hastie, Andreas Buja, and Robert Tibshirani. Penalized Discriminant Analysis. *The Annals of Statistics*, 23(1):73–102, 1995. 14

W.K. Hastings. Monte Carlo samping methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970. 43

David A. Helweg, Douglas H. Cato, Peter F. Jenkins, Claire Garrigue, and Robert D. McCauley. Geographic Variation in South Pacific Humpback Whale Songs. *Behaviour*, 135(1):pp. 1–27, 1998. 97

J. Hérault, C. Jutten, and B. Ans. Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Actes du Xème colloque GRETSI*, pages 1017–1020, 1985. 14

N. Hjort, C. Holmes, P. Muller, and S. G. Waller. *Bayesian Non Parametrics: Principles and practice*. 2010. 2, 6, 57, 61

M. Hosam. CRC Press / Chapman and Hall, 209. 62

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24, 1933. 14

H. Ishwaren and M. Zarepour. Exact and Approximate Representations for the Sum Dirichlet Process. *Canadian Journal of Statistics*, 30:269–283, 2002. 68

T. Jebara. *Discriminative, Generative and Imitative learning*. Phd thesis, Media Laboratory, MIT, 2001. 1, 5

T. Jebara. *Machine Learning: Discriminative and Generative (Kluwer International Series in Engineering and Computer Science)*. Kluwer Academic Publishers, Norwell, MA, USA, 2003. 1, 5

H. Jeffreys. *Theory of Probability*. Oxford, third edition, 1961. 52

Alfons Juan and Enrique Vidal. Bernoulli mixture models for binary images. In *ICPR*, pages 367–370. IEEE Computer Society, 2004. 11

Alfons Juan, José García-Hernández, and Enrique Vidal. Em initialisation for bernoulli mixture learning. In Ana Fred, TerryM. Caelli, RobertP.W. Duin, AurélioC. Campilho, and Dick de Ridder, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 3138 of *Lecture Notes in Computer Science*, pages 635–643. 2004. 11

Robert E. Kass and Adrian E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, June 1995. ISSN 01621459. 7, 28, 50, 52

Sadanori Konishi and Genshiro Kitagawa. *Information criteria and statistical modeling*. Springer series in statistics. Springer, New York, 2008. 27

Sharon X Lee and Geoffrey J McLachlan. Finite mixtures of canonical fundamental skew t-distributions: The unification of the restricted and unrestricted skew t-mixture models . *Statistics and Computing*, page 17, 2015. 124

Sharon X. Lee and GeoffreyJ. McLachlan. On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification*, 7(3): 241–266, 2013. ISSN 1862-5347. 11, 125

Steven M. Lewis and Adrian E. Raftery. Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association*, 92:648–655, 1994. 49, 52

B. G. Lindsay. *Mixture Models: Theory, Geometry and Applications*. NSF-CBMS Conference series in Probability and Statistics, Penn. State University, 1995. 10

Smith A. F. M. and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Royal Statistical Society*, pages 3–23, 1993. 51

Steven N. Maceachern. Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23(3):727–741, 1994. 70

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. 21, 23, 45

J-M Marin, K. Mengersen, and C. P. Robert. Bayesian Modelling and Inference on Mixtures of Distributions. *Bayesian Thinking - Modeling and Computation*, (25):459–507, 2005. 2, 3, 6, 7, 34, 77

Jean-Michel Marin and Christian P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York, 2007. 44

F. H. C. Marriott. Separating Mixtures of Normal Distributions. *Biometrics*, 31(3):767–769, 1975. 11, 34

Itay Mayrose, Nir Friedman, and Tal Pupko. A gamma mixture model better accounts for among site rate heterogeneity. In *ECCB/JBI'05 Proceedings, Fourth European Conference on Computational Biology/Sixth Meeting of the Spanish Bioinformatics Network (Jornadas de BioInformática), Palacio de Congresos, Madrid, Spain, September 28 - October 1, 2005*, page 158, 2005. 11

G. J. McLachlan and K. E. Basford. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, 1988. 1, 5, 10, 18

G. J. McLachlan and D. Peel. *Finite Mixture Models*. New York: Wiley, 2000. 1, 5, 10, 11

Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2. ed edition, 2008. 2, 3, 5, 7, 17, 18, 20, 21

G.J. McLachlan, D. Peel, and R.W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3–4):379 – 388, 2003. Recent Developments in Mixture Model. 11

Paul David McNicholas and Thomas Brendan Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008. 11, 15

L. Medrano, M. Salinas, I. Salas, P. Ladrón de Guevara, A. Aguayo, J. Jacobsen, and C. S. Baker. Sex identification of humpback whales, Megaptera novaeangliae, on the wintering grounds of the Mexican Pacific Ocean. *Canadian Journal of Zoology*, 72(10):1771–1774, 1994. 97

E. Mercado and A. Kuh. Classification of humpback whale vocalizations using a self-organizing neural network. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, volume 2, pages 1584–1589 vol.2, May 1998. 97

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087, 1953. 43

S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993. 43

T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. 1, 5

A. Mkhadri, G. Celeux, and A. Nasroallah. Regularization in discriminant analysis: an overview. *Computational Statistics & Data Analysis*, 23(3): 403–423, January 1997. 14

Fionn Murtagh. The Remarkable Simplicity of Very High Dimensional Data: Application of Model-Based Clustering. *Journal of Classification*, 26(3): 249–277, 2009. 14

Daniel J. Navarro, Thomas L. Griffiths, Mark Steyvers, and Michael D. Lee. Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, 50(2):101–122, April 2006. 2, 6, 61

R. Neal and G. E. Hinton. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. Dordrecht: Kluwer Academic Publishers, 1998. 21

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. 2, 3, 6, 7, 43

Radford M. Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2): 249–265, 2000. 2, 6, 61, 68, 71, 74

Michael A. Newton and Adrian E. Raftery. Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1):3–48, 1994. ISSN 00359246. 51

Jana Novovičová and Antonín Malík. Application of multinomial mixture model to text classification. In FranciscoJosé Perales, AurélioJ.C. Campilho, NicolásPérez de la Blanca, and Alberto Sanfeliu, editors, *Pattern Recognition and Image Analysis*, volume 2652 of *Lecture Notes in Computer Science*, pages 646–653. Springer Berlin Heidelberg, 2003. 11

P. Orbanz and Y. W. Teh. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer, 2010. 2, 6, 61

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. 2, 3, 6, 7, 34, 35, 38, 39, 40, 41, 73

Federica Pace, Frederic Benard, Herve Glotin, Olivier Adam, and Paul White. Subunit definition and analysis for humpback whale call classification. *Applied Acoustics*, 71(11):1107 – 1112, 2010. 98

K. Pearson. Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894. 10

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901. 14

D. Peel and G.J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000. 11

G. Picot, O. Adam, M. Bergounioux, H. Glotin, and F.-X. Mayer. Automatic prosodic clustering of humpback whales song. In *New Trends for Environmental Monitoring Using Passive Systems, 2008*, pages 1–6, Oct 2008. 98

J. Pitman. Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields*, 102(2):145–158, 1995. ISSN 0178-8051. 61, 113

J. Pitman. Combinatorial Stochastic Processes. Technical Report 621, Dept. of Statistics. UC, Berkeley, 2002. 2, 6, 62, 67, 68

Saumyadipta Pyne, Xinli Hu, Kui Wang, Elizabeth Rossin, Tsung-I Lin, Lisa M. Maier, Clare Baecher-Allan, Geoffrey J. McLachlan, Pablo Tamayo, David A. Hafler, Philip L. De Jager, and Jill P. Mesirov. Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences*, 106(21):8519–8524, may 2009. 11

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. 1, 5, 112

Adrian E. Raftery. Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 10, pages 163–187. Chapman & Hall, London, UK, 1996. 7, 49, 50, 52

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. 47

C. Rasmussen. The Infinite Gaussian Mixture Model. *Advances in neuronal Information Processing Systems*, 10:554–560, 2000. 2, 6, 61, 68, 69, 74

Andrea Rau, Gilles Celeux, Marie-Laure Martin-Magniette, and Cathy Maugis-Rabusseau. Clustering high-throughput sequencing data with Poisson mixture models. Research Report RR-7786, Nov 2011. 10

G.M. Reaven and R.G. Miller. An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*, 16(1):17–24, 1979. 92

Richard A. Redner and Homer F. Walker. Mixture Densities, Maximum Likelihood and the Em Algorithm. *SIAM Review*, 26(2):195–239, 1984. 21, 77

Sylvia Richardson and Peter J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components. *Journal of the Royal Statistical Society*, 59(4):731–792, 1997. 34, 35, 36, 38, 39, 43, 73, 77

Christian P. Robert. *The Bayesian choice: a decision-theoretic motivation.* Springer-Verlag, 1994. 34, 35, 38, 43, 44, 61

Donald B. Rubin. Comment on The Calculation of Posterior Distributions by Data Augmentation by M.A. Tanner and W.H. Wong. *Journal of the American Statistical Association*, 82(398):543–546, 1987. 51

A. Samé, C. Ambroise, and G. Govaert. An online classification EM algorithm based on the mixture model. *Statistics and Computing*, 17(3): 209–218, 2007. 17, 18

J. Gershman Samuel and David M. Blei. A tutorial on bayesian nonparametric model. *Journal of Mathematical Psychology*, 56:1–12, 2012. 2, 6, 61, 62, 63, 68, 113

J.L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London, 1997. 43

Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, Cambridge, MA, USA, 1999. 14

G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6: 461–464, 1978. 2, 6, 27, 28

A. J. Scott and M. J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, 27:387–397, 1971. 10

A. J. Scott and M. J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, 37:35–43, 1981. 1, 5, 11, 34

J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994. 62, 63, 66, 116

Hichem Snoussi and Ali Mohammad-Djafari. Penalized maximum likelihood for multivariate gaussian mixture. *Bayesian Inference and Maximum Entropy Methods*, pages 36–46, august 2000. 2, 3, 6, 7, 34, 35, 38, 39, 40, 41

Hichem Snoussi and Ali Mohammad-Djafari. Degeneracy and likelihood penalization in multivariate gaussian mixture models. Technical report, University of Technology of Troyes ISTIT/M2S, 2005. 2, 3, 6, 7, 34, 38, 39, 40, 41

C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103, 1904. 14

M. Stephens. *Bayesian Methods for Mixtures of Normal Distributions*. PhD thesis, University of Oxford, 1997. 2, 3, 6, 7, 34, 35, 36, 38, 43

M. Stephens. Bayesian Analysis of Mixture Models with an Unknown Number of Components – An Alternative to Reversible Jump Methods. *Annals of Statistics*, 28(1):40–74, 2000. 35

Matthew Stephens. Dealing with Multimodal Posteriors and Non-Identifiability in Mixture Models. Technical report, Department of Statistics, Oxford University, 1999. 77

Erik B. Sudderth. *Graphical Models for Visual Object Recognition and Tracking*. PhD thesis, Cambridge, MA, USA, 2006. 71

M. Svensen and C. Bishop. Robust Bayesian mixture modelling. *Neurocomputing*, 64:235–252, 2005. 11

Martin A. Tanner and Wing Hung Wong. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398):528–550, 1987. 44, 49, 51, 74

Yee W. Teh and Michael Jordan. *Hierarchical Bayesian Nonparametric Models with Applications*. Cambridge University Press, Cambridge, UK, 2010. 4, 8, 61, 115, 116

Yee Whye Teh. Dirichlet process. In *Encyclopedia of Machine Learning*, pages 280–287. 2010. 63, 67

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 4, 8, 109, 112, 113, 115, 116, 117, 118

Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, Septembre 2001. 14

Michael E. Tipping and Chris M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999. 14

D. Titterington, A. Smith, and U. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, 1985. 1, 5, 10

J. Van Gael, Y. Saatci, Y.W. Teh, and Z. Ghahramani. Beam sampling for the infinite hidden Markov model. In *Proceedings of the 25th international conference on Machine learning*, pages 1088–1095. ACM New York, NY, USA, 2008. 116, 118

V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 1999. 1, 5

V. N. Vapnik and V. Chervonenkis. Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya (theory of pattern recognition: Statistical problems of learning). Moscow: Nauka, 1974. 1, 5

Isabella Verdinelli and Larry Wasserman. Bayesian analysis of outlier problems using the gibbs sampler. *Statistics and Computing*, 1(2):105–117, 1991. doi: 10.1007/BF01889985. 34

Irene Vrbik and Paul D. McNicholas. Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics & Data Analysis*, 71:196 – 210, 2014. 125

Haixian Wang and Zilan Hu. On em estimation for mixture of multivariate t-distributions. *Neural Processing Letters*, 30(3):243–256, 2009. 11

Larry Wasserman. Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, 44(1):92 – 107, 2000. 50

F. Wood and M. J. Black. A nonparametric Bayesian alternative to spike sorting. *Journal of Neuroscience Methods*, 173(1):1–12, 2008. 61, 68, 69, 74, 116

F. Wood, Thomas L. Griffiths, and Z. Ghahramani. A Non-Parametric Bayesian Method for Inferring Hidden Causes. In *UAI*, 2006. 69, 113

Frank Wood. *Nonparametric Bayesian Models for Neural Data*. PhD thesis, Brown University, 2007. 71

C. F. Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103, 1983. 21

List of Figures

List of Tables

List of Algorithms

List of my publications

Bartcus, M., Chamroukhi, F., Glotin, H. Hierarchical Dirichlet Process Hidden Markov Model for Unsupervised Bioacoustic Analysis. In: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN). Killarney, Ireland, July 2015. 112

Bartcus, M., Chamroukhi, F. Hierarchical Dirichlet Process Hidden Markov Model for unsupervised learning from bioacoustic data. In: Proceedings of the International Conference on Machine Learning (ICML) workshop on unsupervised learning from big bioacoustic data (uLearnBio). Beijing, China, June 2014.

Bartcus, M., Chamroukhi, F., Glotin, H. Clustering Bayésien Parcimonieux Non-Paramétrique. In: Proceedings of 14mes Journées Francophones Extraction et Gestion des Connaissances (EGC), Atelier CluCo: Clustering et Co-clustering. Rennes, France, pp. 3–13, Janvier 2014. 3, 7

Bartcus, M., Chamroukhi, F., Razik, J., Glotin, H. Unsupervised whale song decomposition with Bayesian non-parametric Gaussian mixture. In: Proceedings of the Neural Information Processing Systems (NIPS), workshop on Neural Information Processing Scaled for Bioacoustics: NIPS4B. Nevada, USA, pp. 205–211, December 2013. 3, 7, 99

Chamroukhi, F., Bartcus, M., Glotin, H. Dirichlet Process Parsimonious Mixture for clusteringPreprint, 35 pages, available online arXiv:501.03347. Submitted to Patter Recognition - Elsevier, January 2015. 3, 7

Chamroukhi, F., Bartcus, M., Glotin, H.b. Bayesian Non-Parametric Parsimonious Gaussian Mixture for Clustering. In: Proceedings of 22nd International Conference on Pattern Recognition (ICPR). Stockholm, Sweden, August 2014. 3, 7

Chamroukhi, F., Bartcus, M., Glotin, H.a. Bayesian Non-Parametric Parsimonious Clustering. In: Proceedings of 22nd European Symposium on Artifcial Neural Networks, Computational Intelligence and Machine Learning (ESANN). Bruges, Belgium, April 2014. 3, 7