Clustering Bayésien Parcimonieux Non-Paramétrique

Marius Bartcus*,** Faicel Chamroukhi*,** Hervé Glotin*,**,***

*Université de Toulon, CNRS, LSIS, UMR 7296, 83957 La Garde, France
**Aix Marseille Université, CNRS, ENSAM, LSIS, UMR 7296, 13397 Marseille, France
***Institut Université de France, 75015 Paris, France
{nom}@univ-tln.fr

Résumé. Cet article propose une nouvelle approche Bayésienne non paramétrique pour la classification automatique (clustering). Elle s'appuie sur un modèle de mélange Gaussien infini avec une décomposition en valeurs propres de la matrice de covariance de chaque classe. Le distribution a priori sur les partitions choisie est celle du processus du restaurant chinois (CRP). Cette distribution a priori permet de contrôler la complexité du modèle en reposant sur une formulation statistique solide, et d'estimer automatiquement les nombres de classes à partir des données. De plus, la décomposition en valeurs propres de la matrice de covariance permet d'avoir des modèles flexibles allant du modèle sphérique le plus simple au modèle général qui est plus complexe. L'apprentissage des différents modèles s'effectue par un échantillonnage MCMC de Gibbs. L'approche à été appliquée sur des données simulées et des jeux de données réelles standard afin de valider l'approche et l'évaluer. Les résultats obtenus mettent en évidence l'intérêt du modèle de mélange parcimonieux infini proposé.

1 Introduction

La classification automatique (ou en anglais clustering), est l'une des tâches essentielles en apprentissage automatique et en statistique. L'une des approches les plus populaires en clustering est celle basée sur les modèles de mélange paramétriques finis (McLachlan et Peel., 2000; Fraley et Raftery, 2002; Robert, 2006). Cependant, ces modèles paramétriques se trouvent inadaptés pour représenter des ensembles de données réelles et complexes. L'autre problème de l'approche de clustering à base du modèle du mélange paramétrique fini est celui du choix du nombre de classes, à savoir le problème de sélection de modèle.

Les méthodes Bayésiennes Non Paramétriques (BNP) (Hjort et al., 2010) pour le clustering, y compris le modèle de mélange Gaussien (GMM) infini (Rasmussen, 2000), les CRP et processus de Dirichlet (DP) dans leur version mélange pour le clustering, (Samuel et Blei, 2012; Sudderth, 2006; Pitman, 1995; Aldous, 1985; Ferguson, 1973), fournissent une alternative pertinente pour surmonter ces problèmes. Ils permettent d'éviter de supposer des formes paramétriques restreintes, et permettent ainsi d'inférer la complexité et la structure du modèle à partir des données. L'aspect non-paramétrique de ces approches concerne le fait de supposer que la complexité du modèle associé au nombre de paramètres du modèle croît avec le nombre

et la complexité des données. Ils représentent également une bonne alternative au problème difficile de sélection de modèle rencontrés dans les modèles paramétriques finis. Dans ce travail, nous nous appuyons sur cette formulation bayésienne non paramétrique du mélange Gaussien (GMM) et effectuons une décomposition en valeurs propres de la matrice de covariance de chaque densité Gaussienne comme dans Celeux et Govaert (1995) Banfield et Raftery (1993) pour les GMM finis. Cela conduit à un mélange Gaussien parcimonieux infini qui est plus flexible en termes de modélisation et de son utilisation en clustering, et fournit automatiquement le nombre de classes.

Ce papier est organisé comme suit. La Section 2 rappelle brièvement l'état de l'art sur les mélanges finis parcimonieux pour le clustering. Ensuite, la Section 3 présente la nouvelle approche de clustering parcimonieux non paramétrique proposée. Dans la section 4, on étudie expérimentalement la performance de l'approche proposée en l'appliquant sur des données simulées et réelles.

Notons par $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ un échantillon de n individus i.i.d dans $\mathbb{R}^d$, et soit $\mathbf{z} = (z_1, \ldots, z_n)$ les labels correspondants inconnus où $z_i \in \{1, \ldots, K\}$ représente le label du $i$ème individu $\mathbf{x}_i$, K étant le nombre de classes éventuellement inconnu.

2 Clustering paramétrique Gaussien parcimonieux

Le clustering paramétrique est en général basé sur le modèle de mélange Gaussien fini (GMM) (McLachlan et Peel., 2000). Dans l'approche GMM fini pour le clustering (McLachlan et Peel., 2000; Fraley et Raftery, 2002), les données suivent la densité mélange suivante :

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

ou $\boldsymbol{\theta} = \{\pi_k, \boldsymbol{\theta}_k\}_{k=1}^{K}$ est le vecteur paramètre du GMM qui comprend les proportions du mélange π_k qui sont non-négatives et somment à 1 et $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ sont le vecteur moyenne et la matrice de covariance pour la $k$ième composante Gaussienne du mélange.

Le mélange Gaussien fini parcimonieux (Celeux et Govaert, 1995; Banfield et Raftery, 1993) exploite une décomposition en valeurs propres des matrices des covariances. Ceci fournit une variété de modèles très flexibles. En effet, la décomposition en valeurs propres de la matrice de covariance de chaque composante gaussienne, permet d'avoir des classes ayant différents volumes, formes et orientations (Celeux et Govaert, 1995; Banfield et Raftery, 1993). Cette paramétrisation de la matrice de covariance est de la forme suivante :

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

où λ_k est un scalaire qui représente le volume du cluster k, $\mathbf{D}_k$ est une matrice orthogonale qui représente son orientation et $\mathbf{A}_k$ est une matrice diagonale de déterminant un qui représente sa forme. Cette décomposition conduit à plusieurs modèles flexibles (Celeux et Govaert, 1995) allant de modèles sphériques simples aux modèles généraux plus complexes. Les paramètres $\boldsymbol{\theta}$ du modèle de mélange peuvent être estimés par maximum de vraisemblance (MV) en maximisant la vraisemblance des données observées $p(\mathbf{X}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \, \mathcal{N}_k(\mathbf{x}_i|\theta_k)$ où dans un cadre de maximum a posteriori (MAP) (cadre Bayésien) en maximisant la loi a posteriori des paramètres suivante : $p(\boldsymbol{\theta}|\mathbf{X}) = p(\boldsymbol{\theta})p(\mathbf{X}|\boldsymbol{\theta})$, $p(\boldsymbol{\theta})$ étant la distribution a priori de $\boldsymbol{\theta}$.

L'estimation du maximum de vraisemblance s'effectue en général par l'algorithme Espérance-Maximisation (EM) (Dempster et al., 1977; McLachlan et Krishnan, 1997), où par l'une de ses extensions comme CEM, GEM, etc. L'estimation du maximum a posteriori peut également être effectuée par l'algorithme EM dans le cas de lois conjuguées comme dans Fraley et Raftery (2007). Les techniques d'échantillonnages Markov Chain Monte Carlo (MCMC) peuvent être utilisées aussi pour estimer les paramètres comme dans (Bensmail et al., 1997; Bensmail et Meulman, 2003; Ormoneit et Tresp, 1998). Pour le cas spécifique du GMM fini parcimonieux, plusieurs algorithmes d'apprentissage ont été proposés. Ils s'appuient en majorité sur une estimation par MV via l'algorithme EM ou sur l'une de ses extensions (Banfield et Raftery, 1993; Celeux et Govaert, 1995). On trouve également la version Bayésienne de l'estimation, par l'algorithme EM comme dans Fraley et Raftery (2007), ou par des techniques d'échantillonnages MCMC, notamment l'échantillonneur de Gibbs comme dans Bensmail et al. (1997); Bensmail et Meulman (2003). Cependant, dans l'approche basée sur le mélange GMM fini pour le clustering, le nombre de classes doit être fourni à l'algorithme. L'un des principaux problèmes de cette approche de clustering à base de modèle de mélange fini est donc celui du choix du nombre de composants du mélange (classes) qui correspond au mieux aux données. Le choix du nombre optimal de classes dans le cas du mélange paramétrique Bayésien ou non Bayésien peut être effectuée via des critères de sélection de modèles se basant sur une log-vraisemblance pénalisée, comme BIC (Schwarz, 1978), AIC (Akaike, 1974), etc.

3 Clustering Bayésien non-paramétrique parcimonieux

Les mélanges bayésiens non-paramétriques (BNP) pour le clustering offrent une alternative reposant sur un formalisme statistique solide pour résoudre ce problème de choix du nombre de classes ; le nombre de classes est inféré directement à partir des données (Hjort et al., 2010; Samuel et Blei, 2012; Sudderth, 2006; Rasmussen, 2000), plutôt qu'en s'appuyant à une approche en deux étapes comme pour les mélanges finis. Ce clustering non-paramétrique basé sur les mélanges infinis suppose que les données observées sont crées par un nombre infini des classes, mais seulement un nombre fini d'entre elles a réellement généré les données. Ceci est effectué en posant un processus général comme distribution a priori sur les partitions possibles, ce qui n'est pas restrictif comparé à l'inférence bayésienne classique, et ce de telle manière que seulement un nombre fini de classes sera réellement actif. Une telle distribution a priori peut être celle du processus du restaurant chinois (CRP) (Aldous, 1985; Pitman, 2002; Samuel et Blei, 2012) ou un processus de Dirichlet dans une version mélange pour le clustering (DPM) (Ferguson, 1973; Samuel et Blei, 2012). Plusieurs modèles bayésiens non-paramétriques ont considéré le cas du modèle de mélange Gaussien (GMM) dans sa version générale (non parcimonieuse). On distingue le mélange Gaussien infini (Rasmussen, 2000), la modélisation par mélange de densités et un processus du restaurant chinois (CRP) (Samuel et Blei, 2012), ou le mélange de processus de Dirichlet (DPM) (Antoniak, 1974; Samuel et Blei, 2012). Pour un état de l'art détaillé sur ces approches, le lecteur peut se referer par exemple à ces deux références Samuel et Blei (2012); Sudderth (2006).

Dans l'approche de clustering Bayésien non-paramétrique (BNP) parcimonieux proposée, nous exploitons la décomposition en valeurs propres de la matrice de covariance de chaque classe comme dans Celeux et Govaert (1995) et Banfield et Raftery (1993) pour les GMM finis, et l'intégrant dans un cadre de mélange Gaussien infini. Cela conduit à un mélange Gaussien

infini parcimonieux qui est très flexible en terme de modélisation, et qui permet d'estimer automatiquement le nombre de classes à partir des données. Nous utilisons le processus du restaurant chinois (CRP) comme distribution a priori sur les partitions.

3.1 Processus du restaurant chinois (CRP) et mélange parcimonieux pour le clustering

Le Processus du restaurant chinois (CRP) fournit une distribution sur les partitions infinies des données, qui représente la distribution sur les entiers positifs $1, \ldots, n$. Considérons la distribution jointe suivante sur les labels correspondants des données : $p(z_1, \ldots, z_n) = p(z_1)p(z_2|z_1)p(z_3|z_1, z_2) \ldots p(z_n|z_1, z_2, \ldots, z_{n-1})$. Chaque terme de cette distribution jointe peut être calculé à partir de l'a priori CRP comme suit. Supposons qu'il y a un restaurant avec un nombre infini de tables et dans lequel les clients viennent s'installer dans les tables. Les clients sont sociables, de telle façon que le $i$ème client s'installe à la $k$ième table avec une probabilité proportionnelle au nombre de clients qui y sont déjà installés (n_k), et peut choisir une nouvelle table avec une probabilité proportionnelle à un petit réel positif α représente le paramètre de concentration pour le CRP. Cela peut être formulé comme suit :

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

où K_+ est le nombre de tables pour lesquelles le nombre de clients installés à la $k$ième table est $n_k > 0$, $k \leq K_+$ signifie que k est une table précédemment occupé et $k > K_+$ signifie une nouvelle table à été choisie pour être occupée. A partir de cette distribution, dans un contexte de clustering, on peut donc commencer avec une seule classe et ensuite affecter de nouvelles données à des classes éventuellement nouvelles. En effet, en clustering se basant sur le CRP, les clients correspondent à des données et les tables à des classes. Dans la version mélange CRP pour le clustering, l'a priori du CRP $\text{CRP}(z_1, \ldots, z_{i-1}; \alpha)$ est complété par une densité de paramètres $\boldsymbol{\theta}$ (par exemple dans le cas du GMM une densité Gaussienne multivariée), pour chaque table (classe), et une distribution a priori (G_0) pour les paramètres. Par exemple, dans le cas du mélange Gaussien on peut utiliser des lois conjuguées, à savoir une distribution normale pour chaque moyenne et une distribution inverse-Wishart pour chaque matrice de covariance.

En terme de modèle génératif, cela correspond au processus suivant. Le $i$ème client qui est installé à la table $z_i = k$ choisit un plat (le paramètre $\boldsymbol{\theta}_{z_i}$) de la distribution a priori des plats de la table (cluster). Cela peut se résumer selon le processus génératif suivant.

$$\boldsymbol{\theta}_i \sim G_0 \tag{4}$$

$$z_i \sim \text{CRP}(z_1, \ldots, z_{i-1}; \alpha) \tag{5}$$

$$\mathbf{x}_i \sim p(.|\boldsymbol{\theta}_{z_i})\cdot \tag{6}$$

Selon ce modèle génératif, les paramètres générés $\boldsymbol{\theta}_i$ présentent une propriété de regroupement automatique, c'est-à-dire qu'ils partagent des valeurs répétées avec une probabilité positive et où les valeurs uniques de $\boldsymbol{\theta}_i$ partagées parmi les variables correspondent à des tirages indépendants de la distribution de base G_0 (Ferguson, 1973; Samuel et Blei, 2012). La structure de valeurs partagées définit une partition des entiers de 1 à n, et la distribution de cette partition est un CRP (Ferguson, 1973; Samuel et Blei, 2012). Dans notre mélange Gaussien parcimonieux

infini proposé, les paramètres $\boldsymbol{\theta}_i$ qui comprennent, pour chaque classes, le vecteur moyen et la matrice de covariance décomposée en valeurs propres, permettra ainsi de fournir des classes plus flexibles avec éventuellement différents volumes, formes et orientations. En terme d'interprétation du processus du restaurant chinois, cela peut être vu comme une variabilité et richesse des plats.

3.2 L'échantillonnage MCMC pour l'apprentissage du modèle

Nous avons utilisé un échantillonnage de Gibbs (Rasmussen, 2000; Neal, 1993; Wood et al., 2006; Samuel et Blei, 2012) pour apprendre le modèle parcimonieux non paramétrique Bayésienne proposée.Les distribution a priori utilisée sur les paramètres du modèle dépend du type du modèle parcimonieux considéré. Ainsi, l'échantillonnage des paramètres varie selon le modèle parcimonieux choisi.Nous avons étudié jusqu'à présent sept modèles parcimonieux, couvrant les trois familles du modèle de mélange : la famille générale, la famille diagonale et la famille sphérique. Le tableau 1 présente les modèles considérés et les distribution a priori correspondant à chaque modèle utilisé dans l'échantillonnage de Gibbs.

Decomposition	Type du Modèle	Prior	Appliqué à
$\lambda\mathbf{I}$	Sphérique	$\mathcal{IG}$	λ
$\lambda_k\mathbf{I}$	Sphérique	$\mathcal{IG}$	λ_k
$\lambda\mathbf{B}$	Diagonal	$\mathcal{IG}$	chaque élément de la diagonale de $\lambda\mathbf{B}$
$\lambda_k\mathbf{B}$	Diagonal	$\mathcal{IG}$	chaque élément de la diagonale de $\lambda_k\mathbf{B}$
$\lambda\mathbf{DAD}^T$	Général	$\mathcal{IW}$	$\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$
$\lambda_k\mathbf{DAD}^T$	Général	$\mathcal{IG}$ et $\mathcal{IW}$	λ_k et $\mathbf{\Sigma} = \mathbf{DAD}^T$
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Général	$\mathcal{IW}$	$\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$

TAB. 1: Les GMMs parcimonieux considérés en paramétrant la matrice de covariance et la distribution a priori associée à chaque cas. $\mathcal{I}$ signifie une distribution inverse, $\mathcal{G}$ une distribution Gamma et $\mathcal{W}$ une distribution de Wishart.

Le pseudo-code 1 montre l'algorithme détaillé pour l'échantillonnage de Gibbs dans l'apprentissage des modèles de mélange infini parcimonieux gaussien. Une des étapes principales dans cet algorithme est l'échantillonnage des étiquettes avec la distribution a priori du processus du restaurant chinois (CRP) décrit dans la section 3.1.

Le coût de la méthode est principalement lié à la simulation des étiquettes z_i (et donc au nombres de classes et au nombre d'individus) et des paramètres $\boldsymbol{\theta}_i$ (donc nombre et dimension des paramètres). Plus précisément, la complexité de chaque itération de Gibbs de l'algorithme proposé est proportionnelle à la valeur actuelle du nombre de classes K (K étant estimé automatiquement), et varie donc aléatoirement d'une itération à l'autre, du fait de la distribution sur le nombre de classes. Asymptotiquement, K tend vers $\alpha\log(n)$ quand n tend vers l'infini (Antoniak, 1974). Chaque itération requiert donc $O(\alpha n\log(n))$ opérations pour simuler les étiquettes des classes z_i. La simulation des paramètres (moyennes et matrices de covariances), requiert quant à elle, dans le pire des cas (matrice de covariance pleine) approximativement $O\left(\alpha\log(n)\left(d+\frac{d(d+1)}{2}\right)\right)$ ce qui nous donne une complexité totale de $O\left(\alpha\log(n)\left(N+d+\frac{d(d+1)}{2}\right)\right)$.

Algorithm 1 L'échantillonnage de Gibbs pour l'IPGMM proposé

Entrées :données $\{\mathbf{x}_i\}$, hyper-paramètres et nombre d'échantillons

1: iteration de Gibbs $q \leftarrow 0$
2: hyper-paramètres $\boldsymbol{\alpha}^{(q)}$
3: Commencer avec une classe : $K_+ \leftarrow 1$
4: **for** $i = 1, \ldots, n$ **do**
5: Simuler le label $z_i^{(q)} \sim \text{CRP}(\{z_1, \ldots, z_n\}_{\setminus z_i}; \alpha^{(q)})$
6: Si $z_i^{(q)} = K_+ + 1$ nous avons une nouvelle classe, et on augmente donc K_+ : $K_+ = K_+ + 1$
7: Simuler les paramètres de classe $\boldsymbol{\theta}_i^{(q)}$ selon la distribution *a priori* comme dans le tableau 1.
8: **end for**
9: Simuler les hyper-paramètres $\boldsymbol{\alpha}^{(q)}$
10: $\mathbf{z}^{(q+1)} \leftarrow \mathbf{z}^{(q)}$
11: $\boldsymbol{\alpha}^{(q+1)} \leftarrow \boldsymbol{\alpha}^{(q)}$
12: $q \leftarrow q + 1$

Outputs : $\{\hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}, \hat{K} = K_+\}$

4 Expérimentations

Nous avons effectué des expérimentations sur des données simulées et réelles afin d'évaluer la méthode non paramétrique proposée. A travers ces expérimentations, nous essayons d'abord de souligner la flexibilité de du mélange bayésien non paramétrique parcimonieux proposé, et ce en termes de modélisation, ainsi qu'en terme de son utilisation pour le clustering et la sélection du nombre de classes. Les résultats numériques sont reportés en termes de comparaisons des valeurs de la log-vraisemblance pour les données observés [1], la partition estimée des données, et l'évaluation du nombre de classes estimé. Quand le nombre de classes de la partition estimée est égal au nombre réel de classes, nous calculons également le taux d'erreur de classification. Nous avons comparé le mélange bayésien non paramétrique parcimonieux, avec notamment l'approche bayésienne paramétrique se basant sur le mélange Gaussien fini.

Pour les approches bayésiennes paramétriques (le cas fini), nous avons utilisé l'échantillonnage de Gibbs pour apprendre les paramètres du modèle. Pour chaque ensemble de données simulées, et pour chaque valeur de K, l'échantillonneur de Gibbs est exécuté 10 fois avec différentes initialisations, dans chaque exécution sont générés 2000 échantillons. La solution correspondant à la plus haute probabilité a posteriori est sélectionnée. La selection du nombre de classes est dans ce cas effectuée par le critère AWE (approximate weight of evidence) comme dans Banfield et Raftery (1993).

Pour l'approche bayésienne non paramétrique proposée (IPGMM), nous avons utilisé l'échantillonneur de Gibbs. De même, l'échantillonneur de Gibbs est exécuté dix fois sur chaque jeu de données et la meilleure solution au sens du maximum a posteriori est alors sélectionnée ; le

1. Pour les mélanges infinis, les proportions du mélange sont estimées à partir de la partition obtenue, comme dans Wood et al. (2006)

nombre de classes étant dans ce cas estimé automatiquement au cours de l'échantillonnage de Gibbs.

4.1 Expérimentations sur des données simulées

Nous avons considéré une situation à deux classes pour illustrer l'approche de clustering proposée. Cette situation est la même que dans Celeux et Govaert (1995) et consiste en un échantillon de $n = 500$ observations simulées selon un mélange gaussien à composantes en $\mathbb{R}^2$ et de paramètres $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (0, 0)^T$, $\boldsymbol{\mu}_2 = (3, 0)^T$, $\boldsymbol{\Sigma}_1 = 100\,\mathbf{I}_2$ et $\boldsymbol{\Sigma}_2 = \mathbf{I}_2$. La figure 1 montre les données simulées et les partitions obtenues par trois modèles de l'approche de clustering bayésien non paramétrique parcimonieux proposée.

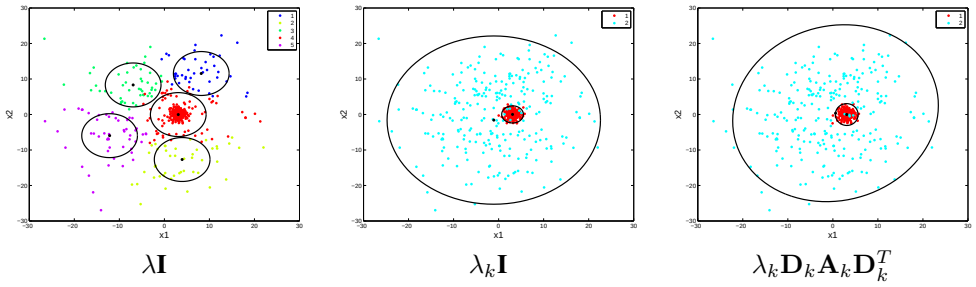

FIG. 1: Un jeu de données à classes et les partition obtenues par le mélange bayésien non-paramétrique pour trois modèles parcimonieux : modèle sphérique à volume égal (gauche), modèle sphérique à volume différent (milieu) et modèle général (droite).

On peut observer que le fait de supposer le même volume pour toutes les classes (modèle $\lambda\mathbf{I}$) ne permet pas de reconstruire la vraie structure cachée des données. Alors que, le modèle parcimonieux dans lequel on suppose seulement que seul le volume des classes peut varier (modèle $\lambda_k\mathbf{I}$), fournit une partition très satisfaisante et qui est très proche de la partition réelle. En effet, le nombre de classes estimé correspond au vrai nombre de classes et le taux d'erreur de classification est de 4.80%. Ce taux d'erreur est même légèrement moins élevé que celui du modèle général (modèle $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$) alors que ce dernier est beaucoup plus complexe. En effet, le modèle le plus général $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ qui permet également d'avoir des volumes de classes différents, fournit un résultat très semblable pour un taux d'erreur est égal à $4,40\%$. Le meilleur modèle au sens de la valeur du log-vraisemblance correspond au modèle sphérique avec différents volumes ($\lambda_k\mathbf{I}$) et surpasse donc même le modèle général qui n'est parcimonieux. On peut conclure que, en ce clustering non-paramétrique basé sur le mélange Gaussien parcimonieux, il est important de prendre en compte des classes avec des volumes différents ; pour cet ensemble de données au moins, le modèle sphérique avec différents volumes ($\lambda_k\mathbf{I}$), est le meilleur modèle.

En termes de comparaison des différents modèles non-paramétriques parcimonieux proposés, et de ces modèles avec l'approche bayésienne paramétrique basée sur le mélange Gaussien fini (GMM), le tableau 2 reporte les valeurs de la log-vraisemblance et du nombre de classes estimé obtenus pour les données simulées par l'approche bayésienne paramétrique basée sur le

mélange Gaussien fini, et l'approche proposée (IPGMM). On peut notamment observer qu'on

TAB. 2: Valeurs de la log-vraisemblance et nombre de classes estimé obtenus pour les données simulées par l'approche bayésienne paramétrique basée sur le mélange Gaussien fini (GMM), et l'approche proposée (IPGMM).

	Données simulées			
	GMM		IPGMM	
Vrai valeur de K	2			
Modèle	$\hat{K}$	log-lik	$\hat{K}$	log-lik
$\lambda\mathbf{I}$	2	-5.5836	5	-5.7707
$\lambda_k\mathbf{I}$	5	-5.1577	2	$\mathbf{-5.1111}$
$\lambda\mathbf{B}$	4	-5.4745	9	-5.4289
$\lambda_k\mathbf{B}$	5	-5.1577	7	-5.2888
$\lambda\mathbf{D}\mathbf{A}\mathbf{D}^T$	2	-5.5608	8	-5.4125
$\lambda_k\mathbf{D}\mathbf{A}\mathbf{D}^T$	4	-5.4175	7	-5.3127
$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	5	$\mathbf{-5.0938}$	2	-5.1194

a deux modèles pour chaque approche pour lesquels le nombre de classes estimé correspond au nombre de classes réel. L'approche proposée peut donc être une bonne alternative moins coûteuse pour l'approche de clustering paramétrique standard.

4.2 Expérimentations sur des données réelles

Nous avons considéré deux jeux de données réels bien connus qui sont Iris et Geyser. Rappelons que Iris est un ensemble de 150 données de dimension 4 comportant trois classes. Le jeux de données Geyser contient 272 individus de dimension 2, le nombre de classes est cependant inconnu mais plusieurs études de clustering le situent entre deux et trois.

La figure 2 montre la partition et les densités estimées par l'approche proposée pour les deux jeux de données pour trois modèles parcimonieux différents dont le modèle général : modèle sphérique (gauche), modèle diagonal (milieu) et modèle général (droite). Les trois modèles permettent d'avoir des classes ayant des volumes différents (à travers λ_k).

Le tableau 3 reporte les résultats numériques pour les deux jeux de données.

Pour le jeu de données iris, on peut remarquer que les deux modèles parcimonieux (sphérique et diagonal, à volume différents) permettent de retrouver le bon nombre de classes et reconstruire la structure des données, alors que le modèle général, qui est plus complexe et est le plus souvent utilisé, quant à lui sous-estime le nombre de classes. Notons que le taux d'erreur pour le modèle diagonal $\lambda_k\mathbf{B}$ est de 5.33% et celui du modèle sphérique $\lambda_k\mathbf{I}$ est de 10.66%. Pour l'approche fini, les modèles pour lesquels on trouve le bon nombre de classes sont les modèles diagonaux $\lambda_k\mathbf{B}$ et $\lambda\mathbf{B}$. Les taux d'erreur correspondant sont respectivement 11.33% $\lambda_k\mathbf{B}$ et 9.33%. Ceci montre un avantage de cette alternative non-paramétrique.

Pour les données Geyser, on peut observer sur les résultats graphiques que les partitions obtenues par l'approche non-paramétrique en utilisant le modèle sphérique à volume différent ($\lambda_k\mathbf{I}$) et le modèle général ($\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$) sont similaires. On peut observer aussi que la partition fournie par le modèle diagonal dans cas infini peut être retenue. Ensuite, on peut noter

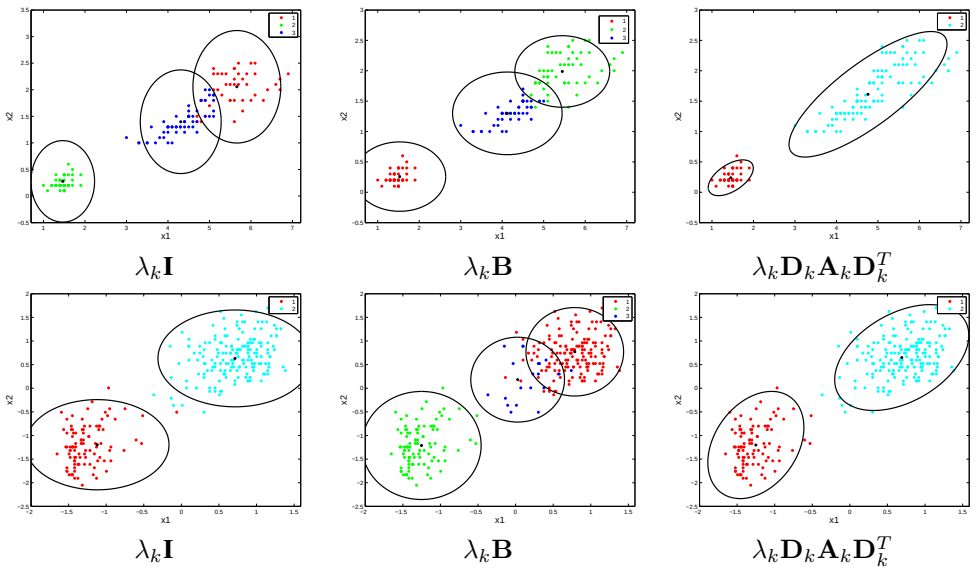

FIG. 2: Résultats obtenus pour les Iris (haut) et Geyser (bas) obtenues par les trois GMMs infinies parcimonieux : modèle sphérique (gauche), diagnoal (milieu) et général (droite).

TAB. 3: Valeurs de la log-vraisemblance et nombre de classes estimé obtenus pour les données simulées par l'approche bayésienne paramétrique basée sur le mélange Gaussien fini (GMM), et l'approche proposée (IPGMM), pour les données Iris et Geyser.

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

également que, sauf pour le modèle sphérique à volume égal ($\lambda\mathbf{I}$), le nombre de classes estimé est conforme à celui de la littérature (entre deux et trois). L'approche paramétrique quant à elle l'estime à deux sauf également pour le cas sphérique à volume égal.

5 Conclusion

Dans cet article, nous avons présenté une nouvelle approche bayésienne non-paramétrique de clustering qui est basée sur un mélange Gaussien infini parcimonieux. Le mélange Gaussien infini parcimonieux se base sur une décomposition en valeurs propres de la matrice de covariance de chaque classe, et le processus du restaurant chinois (CRP) comme a priori. Cette approche permet de dériver plusieurs modèles flexibles et évite le problème difficile de sélection de modèle rencontré dans l'approche paramétrique des mélanges Gaussiens. Nous avons illustré cette méthode sur des données simulées et nous l'avons appliqué sur des données réelles. Les résultats obtenus mettent en évidence l'intérêt d'utiliser le clustering bayésien non-paramétrique parcimonieux comme une bonne alternative pour le clustering parcimonieux à base de GMM fini. Notre travail actuel portent sur plus d'expériences sur des données réelles et simulées. Un des points qui reste ouvert comme en toute approche de clustering est celui de l'évaluation de la partition obtenue. Les futurs travaux concerneront également d'autres techniques pour apprendre les modèles, notamment des méthodes variationnelles.

Références

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

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été St Flour 1983*, pp. 1–198. Springer-Verlag. Lecture Notes in Math. 1117.

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

Banfield, J. D. et A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics 49*(3), 803–821.

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

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

Celeux, G. et G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition 28*(5), 781–793.

Dempster, A. P., N. M. Laird, et D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of The Royal Statistical Society, B 39(1)*, 1–38.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Fraley, C. et A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97*, 611–631.

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

Hjort, N., H. C., P. Muller, et S. G. Waller (2010). *Bayesian Non Parametrics*. Cambrige University Press.

McLachlan, G. J. et T. Krishnan (1997). *The EM algorithm and extensions*. New York : Wiley.

McLachlan, G. J. et D. Peel. (2000). *Finite mixture models*. New York : Wiley.

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

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

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields 102*(2), 145–158.

Pitman, J. (2002). Combinatorial stochastic processes. Technical Report 621, Dept. of Statistics. UC, Berkeley.

Rasmussen, C. (2000). The infinite gaussian mixture model. *Advances in neuronal Information Processing Systems 10*, 554 – 560.

Robert, C. (2006). *Le choix bayésien : principes et pratique*. Statistique et probabilités appliquées. Springer.

Samuel, J. G. et D. M. Blei (2012). A tutorial on bayesian non-parametric model. *Journal of Mathematical Psychology 56*, 1–12.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*, 461–464.

Sudderth, E. B. (2006). *Graphical models for visual object recognition and tracking*. Ph. D. thesis, Massachusetts Institute of Technology, Cambridge, MA, USA.

Wood, F., T. L. Griffiths, et Z. Ghahramani (2006). A non-parametric bayesian method for inferring hidden causes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Summary

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