# Clustering Bayésien Parcimonieux Non-Paramétrique

Marius Bartcus Faicel Chamroukhi Hervé Glotin

Université du Sud Toulon Var
nom@univ-tln.fr

Janvier 28, 2014

# Plan

1. Modèle du mélange Gaussien fini (GMM) pour le clustering
2. Modèle de mélange Gaussien fini parcimonieux
3. Modèle de mélange Gaussien infini parcimonieux (IPGMM)
    - Processus du restaurant chinois (CRP)
    - Le clustering proposé avec CRP et GMM parcimonieux
4. Conclusion et perspectives

# Modèle de mélange fini

## Definition

Densité mélange:

$$f(X,\boldsymbol{\theta}) = \sum_{k=1}^{K} p(z_i = k) f(\mathbf{x}_i | z_i = k; \boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k f_k(X; \boldsymbol{\theta}_k) \quad (1)$$

1. $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)$ - les données observées
2. $\mathbf{Z} = \{1..K\}$ - les données cachées $p(z_i = k) = \pi_k$
3. $(\mathbf{X}, \mathbf{Z})$ - les données complète
4. $f_k(\mathbf{x}_i | \boldsymbol{\theta}_k)$ - loi de probabilité avec les paramètres $\boldsymbol{\theta}_k$
5. $\pi_k$ la probabilité pour le $k^{ime}$ composant
6. $K$ - les nombre des composants dans le mélange

$f_k(\mathbf{x}_i | \boldsymbol{\theta}_k) = \mathcal{N}_k(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ pur les modèle de mélange Gaussien

# Modèle de mélange Gaussien fini

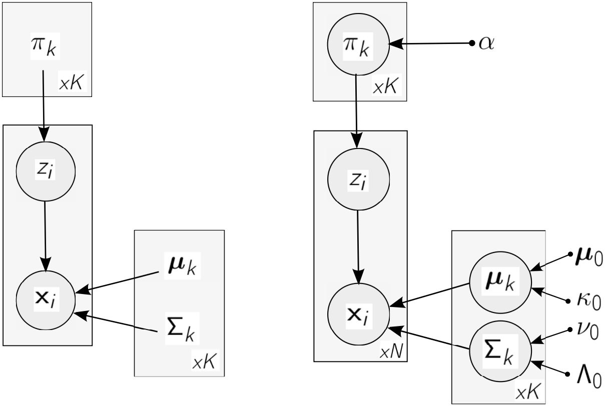



Figure : Les modèles graphiques: le modèle de mélange gaussien fini (GMM) à gauche et le GMM bayésienne à droite

- $\mathcal{H} = \mu_0, \kappa_0, \nu_0, \Lambda_0$
- $\mu_0$ - mean of the data, $\kappa_0$ - shrinkage, $\nu_0$ - degrees of freedom, $\Lambda_0$ - matrix called scale of $\mathcal{IW}$ prior

# Estimation des parameters $\boldsymbol{\theta}$.

| estimateur du MV | estimateur du MAP |
|---|---|
| $L_{ML} = \log p(X\|\boldsymbol{\theta})$<br>$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} L_{ML}(\boldsymbol{\theta}\|X)$ | $L_{MAP}(\boldsymbol{\theta}\|X) = \log p(\boldsymbol{\theta}\|X)$<br>$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}}(\log p(X\|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))$ |

ou $p(\mathbf{X}|\boldsymbol{\theta})$ est la vraisemblance et $p(\boldsymbol{\theta})$ est la distribution priori des paramètres $\boldsymbol{\theta}$.

- La vraisemblance:

$$\mathcal{L}(\theta;X)=\prod_{i=1}^{N} p(x_i;\theta)=\prod_{i=1}^{N}\sum_{k=1}^{K} f_k(x_i;\theta_k) \tag{2}$$

- La vraisemblance de données observées:

$$\log \mathcal{L}(\theta;X)=\log\prod_{i=1}^{N} p(x_i;\theta)=\sum_{i=1}^{N}\log\sum_{k=1}^{K}\pi_k f_k(x_i;\theta_k) \tag{3}$$

- La vraisemblance de données complète:

$$\begin{aligned}\log \mathcal{L}_c &= \log\prod_{i=1}^{N} p(x_i,z_i;\theta)=\sum_{i=1}^{N}\log\prod_{k=1}^{K}[p(z_i=k)p(x|z_i=k;\theta_k)]^{z_{ik}} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\log\pi_k f_k(x_i;\theta_k)\end{aligned} \tag{4}$$

  ou $z_{ik} = 1$ si $z_i = k$ sinon $z_{ik} = 0$.

L'estimation des paramètres s'effectue par:

- EM ou une de extensions comme CEM, GEM, etc. par MV ou MAP
- Les méthodes MCMC: L'échantillonnage de Gibbs pour le cas bayésien

# MCMC: L'échantillonnage de Gibbs

**Algorithm 1** L'échantillonnage de Gibbs pour le modèle de mélange fini

**Entrées:** Les données $\mathbf{x}_i$, nombre de clusters $K$, nombre d'échantillons $n_s$.
Initialisation de $\boldsymbol{\pi}^{(0)}$ et $\boldsymbol{\theta}^{(0)}$
**for** $q = 1$ **to** $n_s$ **do**
  **for** $k = 1$ **to** $K$ **do**
    Évaluer les probabilités postérieures $\tau_{ik}^{(q)} = \frac{\pi_k^{(q-1)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q-1)})}{\sum_{k=1}^{K} \pi_k^{(q-1)} f_k(\mathbf{x}_i|\boldsymbol{\theta}_k^{(q-1)})}$
    Simuler $\boldsymbol{\pi}_k^{(q)}$ à partir de $p(\boldsymbol{\pi}_k|\tau_{ik}, \boldsymbol{\theta}_k, \mathbf{X})$
    Simuler $\boldsymbol{\theta}_k^{(q)}$ à partir de $p(\boldsymbol{\theta}_k|\tau_{ik}, \boldsymbol{\pi}_k, \mathbf{X})$
  **end for**
**end for**

Pour GMM:

- $f_k(\mathbf{x}_i|\boldsymbol{\theta}_k) = \mathcal{N}_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- $p(\boldsymbol{\pi}_k|\tau_{ik}, \boldsymbol{\theta}_k, \mathbf{X}) \sim \mathcal{D}ir(\boldsymbol{\alpha})$
- $p(\boldsymbol{\theta}_k|\tau_{ik}, \boldsymbol{\pi}_k, \mathbf{X}) = p(\boldsymbol{\mu}_k|\Sigma_k, \tau_{ik}, \boldsymbol{\pi}_k, \mathbf{X}) p(\Sigma_k|\boldsymbol{\mu}_k, \tau_{ik}, \boldsymbol{\pi}_k, \mathbf{X}) \sim \mathcal{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \Lambda_0)$

# La sélection du modèle pour le cas fini ($K = ?$)

$$e.g \quad \text{maximiser} \quad BIC = \log \mathcal{L}(\boldsymbol{\theta}_k; X)_M - \frac{\nu_M \log(n)}{2}$$

- $\mathcal{L}(\boldsymbol{\theta}_k; X)_M$ - la log vraisemblance maximisé pour la modèle $M$
- $\nu_M$ - le nombre independent des paramètres à estimé dans la modèle $M$
- $n$ - le nombre d'observation.

**Algorithm 2** Classification des données dans le cadre des modèles de mélange fini

1. Fixé $K_{max}$
2. Lancé $K_{max}$ fois l'algorithme d'apprentissage(EM/Gibbs) et choisir le meilleur modèle.
3. Lancé l'algorithme d'apprentissage et classifié les données

# Parsimonious Gaussian Mixture Model (décomposition en valeur propre)

[Banfield and Raftery(1993)] et [Celeux and Govaert(1995)] exploit une décomposition en valeur propre des matrices des covariance.

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k \tag{5}$$

$\lambda_k$-volumes, $\mathbf{D}_k$-orientations, $\mathbf{A}_k$-formes.

| Decomposition | Type du Modèle | Prior | Appliqué à |
|---|---|---|---|
| $\lambda\mathbf{I}$ | Sphérique | $\mathcal{IG}$ | $\lambda$ |
| $\lambda_k\mathbf{I}$ | Sphérique | $\mathcal{IG}$ | $\lambda_k$ |
| $\lambda\mathbf{B}$ | Diagonal | $\mathcal{IG}$ | $diag(\lambda\mathbf{B})$ |
| $\lambda_k\mathbf{B}$ | Diagonal | $\mathcal{IG}$ | $diag(\lambda_k\mathbf{B})$ |
| $\lambda\mathbf{DAD}^T$ | Général | $\mathcal{IW}$ | $\mathbf{\Sigma} = \lambda\mathbf{DAD}^T$ |
| $\lambda_k\mathbf{DAD}^T$ | Général | $\mathcal{IG}$ et $\mathcal{IW}$ | $\lambda_k$ et $\mathbf{\Sigma} = \mathbf{DAD}^T$ |
| $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ | Général | $\mathcal{IW}$ | $\mathbf{\Sigma}_k = \lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ |

Table : 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. $diag(.)$ signifie chaque element de la diagonale d'une matrice.

# Modèle de mélange Gaussien infini parcimonieux (IPGMM)

- IGMM proposé par [Rasmussen(2000)].
- Processus du restaurant chinois (CRP). $K \to \infty$
- Estimation par MAP MCMC (L'échantillonnage de Gibbs).

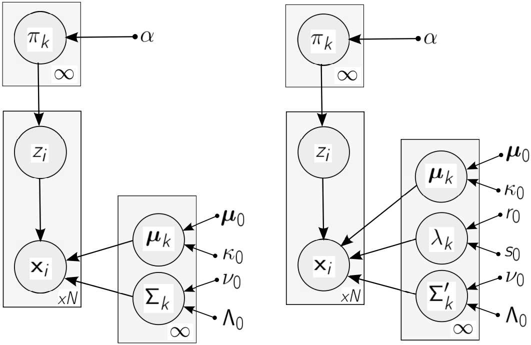



Figure : Les modèles graphiques pour le mélange gaussien infini (IGMM) à gauche et le mélange gaussien infini parcimonieux proposé à droit (IPGMM)

# Processus du restaurant chinois (CRP)

Supposons qu'il y a un restaurant avec un nombre infini des tables et dans lequel les clients viennent s'installer dans les tables.


1. Le premiere client s'installe à la 1ère table
2. Le deuxième client s'installe à la 1ère table avec une probabilité $\frac{1}{1+\alpha}$ où à la 2ème table avec la probabilité $\frac{\alpha}{1+\alpha}$
3. ...
4. Le *n*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 $\alpha$ représente le paramètre de concentration pour le CRP.

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

où $K_+$ - nombre de tables avec $n_k > 0$, $k \leq K_+$ signifie que $k$ est une table précédemment occupé et $k > K_+$ signifie une nouvelle table à été choisie pour être occupée.

# Le clustering proposé avec CRP et GMM parcimonieux

**Algorithm 3** L'échantillonnage de Gibbs pour l'IPGMM proposé

**Entrées :** les données $x_i$, les hyper-paramètres $\mathcal{H}$ le nombre d'échantillons $n_s$
Initialisation des labels $z_1 \dots z_n \leftarrow 0$, $\mathbf{Z}_0 \leftarrow \{z_1, \dots z_n\}$ et les nombre active des clusters $K_+ \leftarrow 1$.
**for** $q = 1$ **to** $n_s$ **do**
  $\mathbf{Z}^{(q)} \leftarrow \mathbf{Z}^{(q-1)}$
  $\boldsymbol{\theta}^{(q)} \leftarrow \boldsymbol{\theta}^{(q-1)}$
  **for** $i = 1, \dots, n$ **do**
    **if** $z_i \neq 0$ **then**
      $n_k = \sum_{j=1}^{n} p(z_i = k)$
      **if** $n_k - 1 = 0$ **then**
        $\boldsymbol{\theta}^{(q)} \leftarrow \boldsymbol{\theta}^{(q)} \backslash \boldsymbol{\theta}_{(z_i)}$
        $z_j = z_j - 1 \forall j > i$
        $K_+ = K_+ - 1$
      **end if**
    **end if**
    Simuler le label $z_i \sim \text{CRP}(\{z_1, \dots, z_n\}_{\backslash z_i}; \alpha^{(q)})$
    **if** $z_i \in \mathbf{Z}^{(q)}$ **then**
      Simuler les paramètres de classe $\boldsymbol{\theta}_{z_i}^{(q)}$ selon le posterior comme dans le tableau
      $\boldsymbol{\theta}^{(q)} \leftarrow \{\boldsymbol{\theta}^{(q-1)}, \{\boldsymbol{\mu}_{z_i}, \Sigma_{z_i}\}\}$
      Nous avons une nouvelle classe, et on augmente donc $K_+$ : $K_+ = K_+ + 1$
    **end if**
  **end for**
  **for** $i = 1, \dots, K_+$ **do**
    Simulé les paramètres du modèle $\boldsymbol{\theta}_k^{(q)}$ selon la distribution priori comme dans le tableau
  **end for**
**end for**

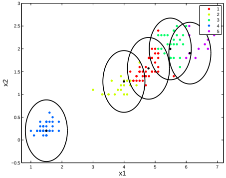
# Experimentations sur les données Iris

- Lancement de notre algorithmes 100 fois
- Affichage de taux d'erreur et l'écart-type

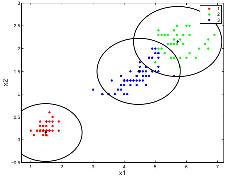
| Modèle | GMM | Trouve le vrai # des classes | taux d'erreur |
|---|---|---|---|
| $\lambda\mathbf{B}$ | fini | 100% | $10\% \pm 0.21\%$ |
| $\lambda_k\mathbf{B}$ | fini | 100% | $10\% \pm 2.81\%$ |
| $\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$ | infini | 100% | $30.66\% \pm 14.14\%$ |
| $\lambda_k\mathbf{DAD}^T$ | infini | 85% | $2.66\% \pm 1.47\%$ |
| $\lambda_k\mathbf{B}$ | infini | 79% | $4\% \pm 0.45\%$ |
| $\lambda\mathbf{I}_d$ | infini | 97% | $11.33\% \pm 1.15\%$ |

Table : Résultats obtenus pour les Iris pour le cas de GMMs fini et GMMs infini.

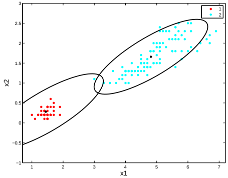
# Experimentations sur les données Iris



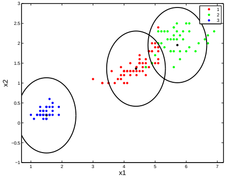

$\lambda\mathbf{I}$: $\hat{K} = 5, \hat{\mathcal{L}} = -1.6180$



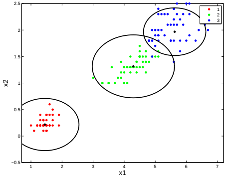

$\lambda_k\mathbf{B}$: $\hat{K} = 3, \hat{\mathcal{L}} = -1.7041$



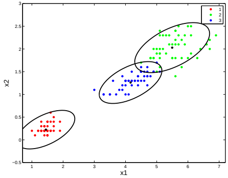

$\lambda_k\mathbf{DAD}^T$: $\hat{K} = 2, \hat{\mathcal{L}} = -1.6202$




$\lambda\mathbf{I}$: $\hat{K} = 3, \hat{\mathcal{L}} = -1.5799$




$\lambda_k\mathbf{B}$: $\hat{K} = 3, \hat{\mathcal{L}} = -1.5805$




$\lambda_k\mathbf{DAD}^T$: $\hat{K} = 3, \hat{\mathcal{L}} = -1.5583$

Figure : Résultats obtenus pour les Iris pour le cas de GMMs fini (haut) et GMMs infini (bas) obtenues par les trois modèles parcimonieux : sphérique (gauche), diagnoal (milieu) et général (droite).

# Conclusion et Perspectives

Conclusion:

- Nous avons proposé une nouvelle approche bayésienne non-paramétrique
- Modélisation parcimonieuse flexible
- Premieres résultats encourageants

Perspectives:

- Expériences sur d'autre données réelles standard et des données bioacoustique
- Étudier d'autre modèle parcimonieuse pour IPGMM
- Étudier d'autre techniques MCMC

Merci!

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

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

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