Approche bayésienne de clustering sur le jeu de données des galaxies.

Bastien Maudry, Nino Scheele
Encadré par Martin Metodiev

Polytech Clermont

18.02.2025

Introduction

L'approche bayésienne de clustering rejoint les algorithmes d'apprentissage non supervisé, particulièrement utiles pour trouver du sens dans des jeux de données complexes. Ils ont des applications dans de nombreux domaines.
Le jeu de données des galaxies regroupe les vélocités de 82 galaxies, il est couramment utilisé dans la recherche pour étudier le clustering.

Plan

Plan

Fondement théorique : Modèle de mélange gaussien

Soient $y = (y_1, \ldots, y_N)$ nos observations suivant une loi de vraisemblance :

$$p(y_1, \ldots, y_N \mid \theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k \times \mathcal{N}(y_i; \mu_k, \sigma_k^2),$$

- K le nombre de groupes,
- $\mathcal{N}(y_i; \mu_k, \sigma_k^2)$ la densité d'une loi gaussienne en y_i,
- μ_k la moyenne du groupe k, σ_k^2 sa variance et $\sum_{k=1}^{K} \pi_k = 1$ (avec les $\pi_k > 0$).

Objectif : Déterminer les paramètres $(\pi_1, ..., \pi_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2) = \theta$ du modèle qui décrivent au mieux les observations y.

Fondement théorique : Modèle de mélange gaussien

Pour chaque observation y_i dans un groupe k, on introduit une variable latente Z_i qui correspond à un vecteur dans lequel toutes les composantes sont nulles sauf une qui correspond au groupe auquel appartient y_i. Ainsi,

$$\forall i \in [\![1; N]\!] \; Z_i|\theta \; \sim \mathcal{M}(1, (\pi_1, ..., \pi_K)) \tag{1}$$

Et nous construisons donc la matrice $Z \in \mathcal{M}_{N\times K}(\mathbb{R})$ contenant tous les Z_i. Dans ce cas,

$$\forall i \in [\![1; N]\!], \; y_i|Z_{ik} = 1 \; \sim \mathcal{N}(\mu_k, \sigma_k^2) \tag{2}$$

Remarque : On pose $\mu = (\mu_1, \ldots, \mu_K)$, $\sigma^2 = (\sigma_1^2, \ldots, \sigma_K^2)$ et $\pi = (\pi_1, \ldots, \pi_K)$.

Fondement théorique : Principe bayésien

La loi de Bayes s'écrit dans le cas de notre densité ainsi :

$$p(\theta \mid y_1, \ldots, y_N) = \frac{p(y_1, \ldots, y_N \mid \theta) p(\theta)}{p(y_1, ..., y_N)}$$

- $p(\theta \mid y_1, \ldots, y_N)$ est la densité de la loi a posteriori
- $p(y_1, \ldots, y_N \mid \theta)$ désigne la vraissemblance de nos données
- $p(\theta)$ représente la densité de nos loi a priori

Objectif : maximiser $p(\theta \mid y_1, \ldots, y_N)$, ainsi nous prendrons comme estimateur :

$$\begin{aligned} \hat{\theta}_{MAP} &= \underset{\theta \in \Theta}{\arg\max} \{p(\theta \mid y_1, \ldots, y_N)\} \\ &= \underset{\theta \in \Theta}{\arg\max} \{\log(p(y_1, \ldots, y_N \mid \theta)) + \log(p(\theta))\} \end{aligned}$$

Fondement théorique : Gibbs sampling

Selon la définition donnée par Robert et al. (2007), pour une distribution conjointe $p(\theta_1, \ldots, \theta_n)$ avec leurs lois conditionnelles respectives $p_1, \ldots, p_n$, l'algorithme général du Gibbs Sampling s'écrit, pour $(\theta_1^{(t)}, \ldots, \theta_n^{(t)})$ fixé à l'itération t, comme suit :

- $\theta_1^{(t+1)} \sim p_1(\theta_1 \mid \theta_2^{(t)}, \ldots, \theta_n^{(t)})$;
- $\theta_2^{(t+1)} \sim p_2(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_n^{(t)})$;

  ...
- $\theta_n^{(t+1)} \sim p_n(\theta_n \mid \theta_1^{(t+1)}, \ldots, \theta_{n-1}^{(t+1)})$

Objectif : Découper le problème de manière à calculer $\theta^{(t+1)}$ à partir de $\theta^{(t)}$ en utilisant les probabilités conditionnelles.

Plan

Méthodologie : Choix des a priori

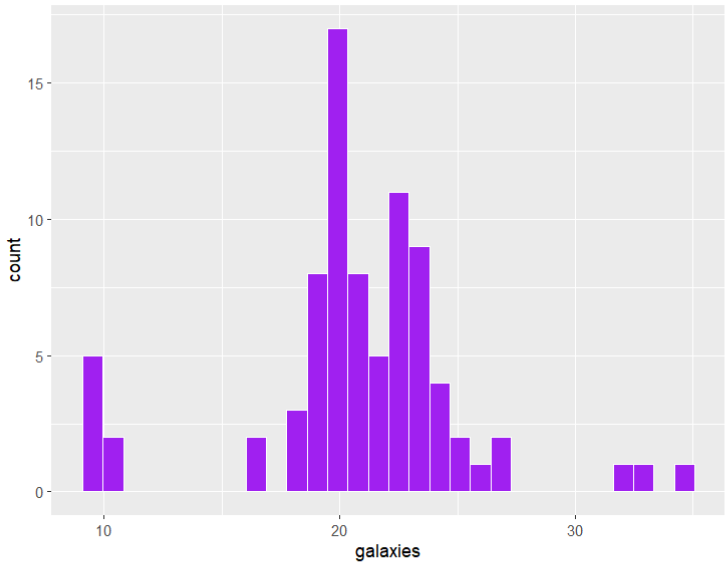
Nos différentes lectures (Aitkin, 2001), (Grün et al., 2022) nous ont permis d'établir que ces familles de lois étaient recommandées :

$$\pi = (\pi_1, ..., \pi_K) \sim Dir(\alpha_1, \ldots, \alpha_K), \quad \alpha_1, \ldots, \alpha_K > 0$$
$$\sigma_k^2 \sim \mathcal{IG}(\nu_0\ , \Lambda_0), \quad \nu_0, \Lambda_0 > 0$$
$$\mu_k | \sigma_k^2 \sim \mathcal{N}(\mu_0\ , \sigma_k^2/\kappa_0), \quad \kappa_0, \mu_0 > 0$$

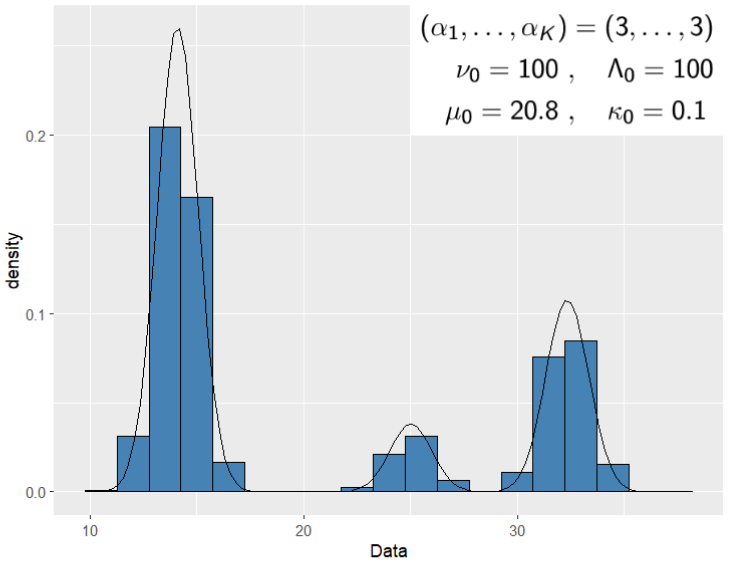
Avec la méthode du prior detective check décrite dans Conn et al. (2018) on choisit alors :

$$(\alpha_1, \ldots, \alpha_K) = (3, \ldots, 3)$$
$$\nu_0 = 100\ , \quad \Lambda_0 = 100$$
$$\mu_0 = 20.8 \quad \text{et} \quad \kappa_0 = 0.1$$

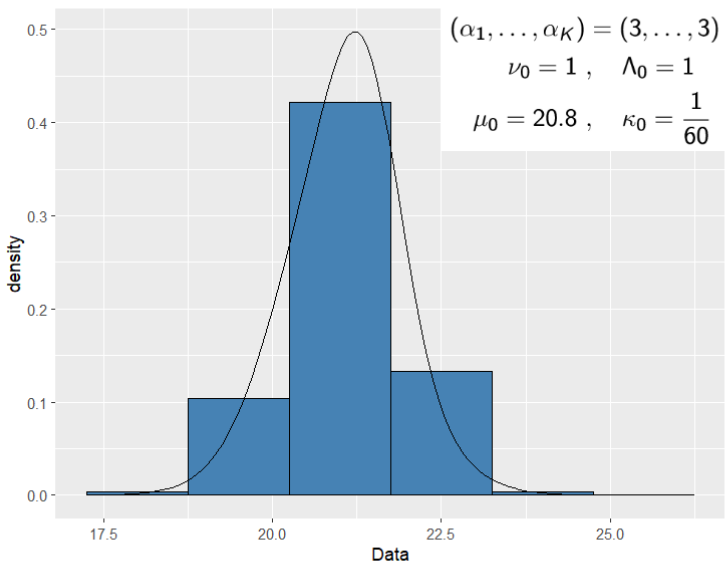
Méthodologie : choix des a priori



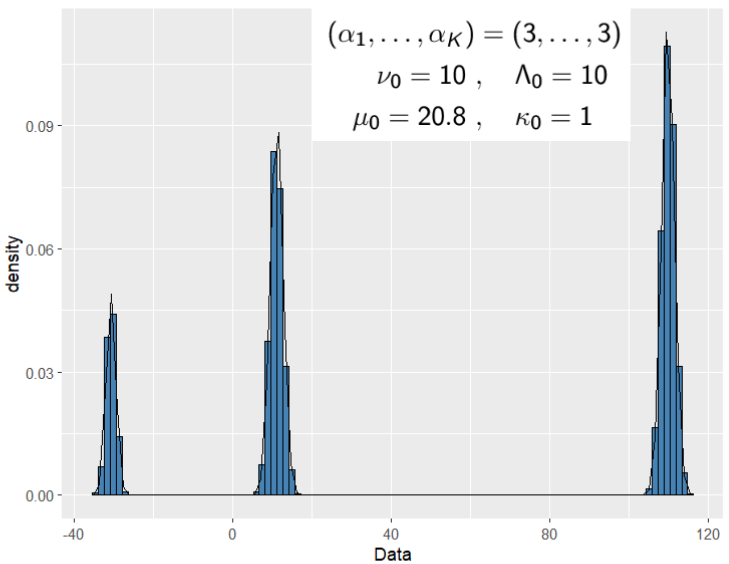
Répartition des galaxies



Répartition de données générées avec les a priori



Répartition de données générées avec les a priori



Répartition de données générées avec les a priori

Méthodologie : Implémentation du Gibbs sampling

Dans le cadre de notre projet, nous implémentons le Gibbs sampling en s'inspirant de Robert et al. (2007) et de Gelman et al. (1995) avec les tirages suivants :

$$\pi \mid y, \mathbf{z} \sim Dir\left(\alpha_1 + m_1(\mathbf{z}),\ \ldots,\ \alpha_K + m_K(\mathbf{z})\right)$$

$$\sigma_k^2 \mid y, \mathbf{z} \sim \mathcal{IG}\left(\frac{\nu_0 + m_k(\mathbf{z})}{2},\ \frac{1}{2}\left[\Lambda_0 + \hat{s}_k^2(y, \mathbf{z}) + \frac{\kappa_0\ m_k(\mathbf{z})}{\kappa_0 + m_k(\mathbf{z})}\right]\right)$$

$$\mu_k \mid y, \mathbf{z}, \sigma_k^2 \sim \mathcal{N}\left(\xi_k(y, \mathbf{z}),\ \frac{\sigma_k^2}{\kappa_0 + m_k(\mathbf{z})}\right)$$

- $m_k(\mathbf{z})$: le nombre d'observations dans le cluster k,
- $\bar{y}_k(\mathbf{z})$: la valeur moyenne des observations du cluster k,
- $\hat{s}_k^2(y, \mathbf{z})$: l'estimation de la variance dans le cluster k,
- $\xi_k(X, \mathbf{z})$;

$$\xi_k(y, \mathbf{z}) = \frac{\kappa_0\ \mu_0 + m_k(\mathbf{z})\ \bar{y}_k(\mathbf{z})}{\kappa_0 + m_k(\mathbf{z})}$$

Méthodologie : Implémentation du Gibbs sampling

Enfin on tire $Z \mid \theta$: (On note $\theta_k = (\pi_k, \mu_k, \sigma_k^2)$)

$$1 \leq i \leq n,\ Z_i \mid y_i, \theta \sim \mathcal{M}_K(1;\ z_{i1},\ \ldots, z_{iK})$$

Avec pour $1 \leq k \leq K$:

$$z_{ik} = \frac{\pi_k \times \mathcal{N}(y_i;\ \mu_k,\ \sigma_k^2)}{\sum_{p=1}^{K}(\pi_p \times \mathcal{N}(y_i;\ \mu_p,\ \sigma_p^2))}$$

Méthodologie : Algorithme du Gibbs sampling

On initialise le Gibbs sampling avec $\mathbf{z}^{(0)}$ généré selon un clustering ascendant hiérarchique.

Algorithm 1 Gibbs sampling

Entrées : $\mathbf{z}^{(0)}$

Pour t allant de 1 à T **:** (T le nombre d'itérations)

Tirer $(\sigma^2)^{(t)}$ avec $\mathbf{z}^{(t-1)}$

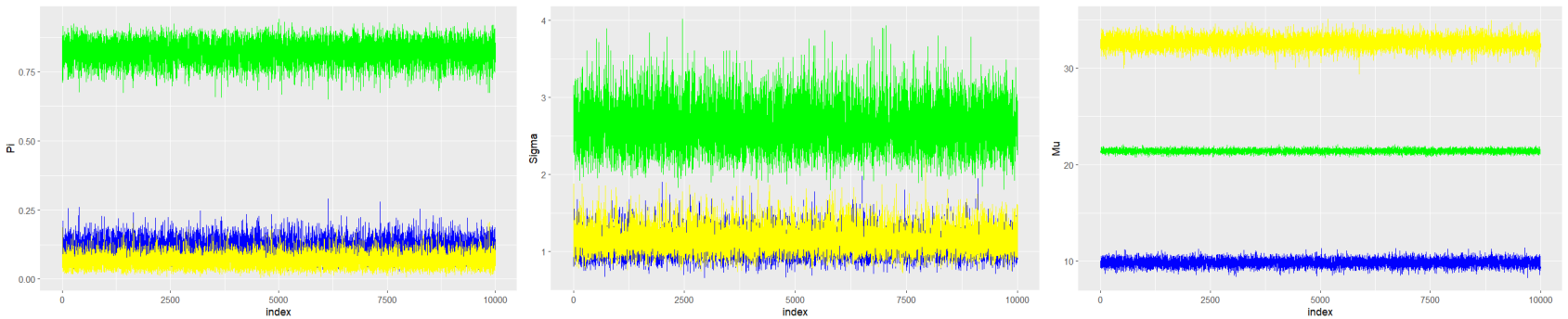
Tirer $\mu^{(t)}$ avec $\mathbf{z}^{(t-1)}, (\sigma^2)^{(t)}$

Tirer $\pi^{(t)}$ avec $\mathbf{z}^{(t-1)}$

Tirer $\mathbf{z}^{(t)}$ avec $\pi^{(t)}$, $\mu^{(t)}$ et $(\sigma^2)^{(t)}$

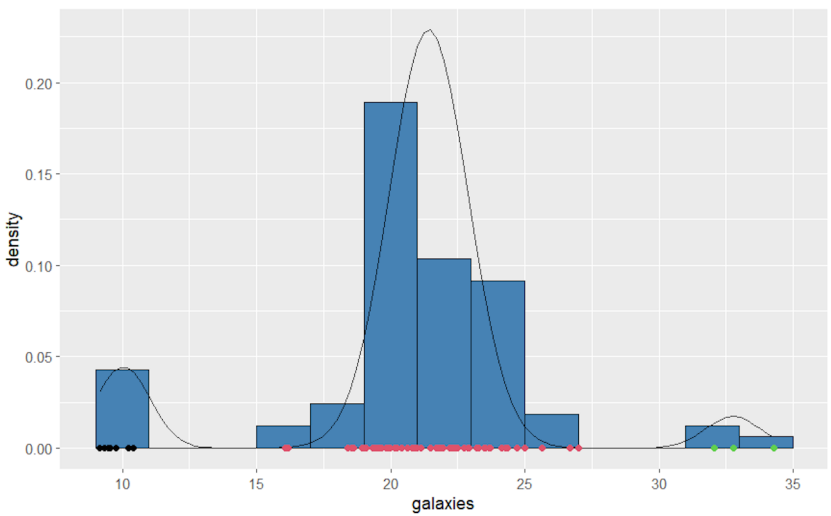
Fin Pour

Méthodologie : 1er résultats



Graphe de sortie de l'échantillonneur de Gibbs après Label switching

$$\hat{\pi} = (0.104, 0.827, 0.0677)$$
$$\hat{\mu} = (9.979, 21.455, 33.048)$$
$$\hat{\sigma}^2 = (1.041, 1.964, 1.098)$$



Visualisation du clustering après calcul du MAP

Méthodologie : Détermination du nombre de clusters

Il existe deux indices pour déterminer le nombre de clusters :

- Le BIC *(Bayesian information criterion)* proposé dans Bouveyron et al. (2019) défini comme suit :

$$BIC_K \;=\; 2\log(\; p(y|\hat{\theta}_{MAP}, K)\;) - \omega_K \log(N)$$

  Avec ω_K le nombre de paramètres libres dans le modèle

- L'ICL *(Integrated Completed Likelihood)* qui prend en compte l'estimation de Z :

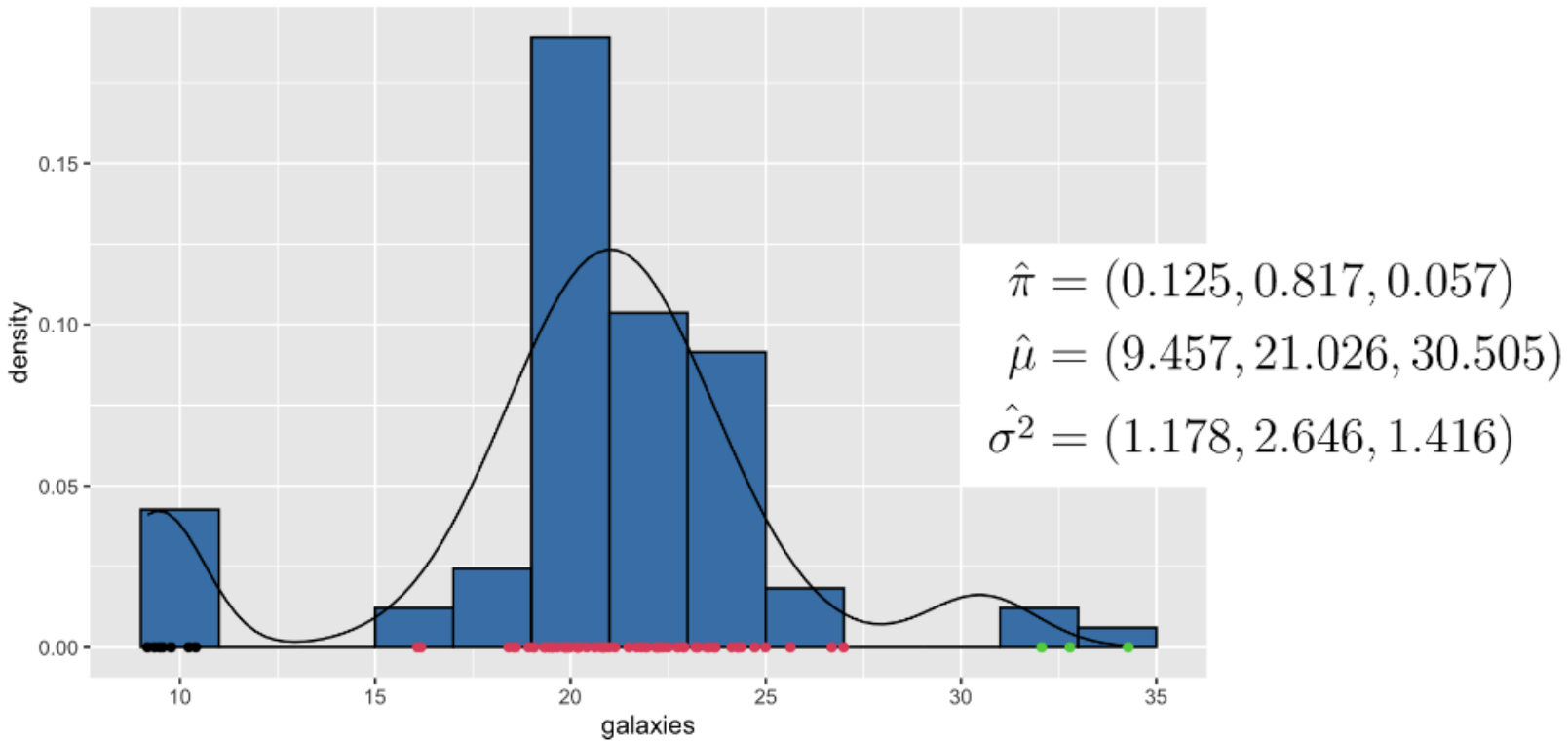
$$ICL_K = BIC_K \;-\; E(K)$$

  Avec $E(K) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \hat{z}_{ik}\log(\hat{z}_{ik})$

Méthodologie : Résultats BIC et ICL

Nombre de groupe	2	3	4	5	6
BIC	-507.9	-468.5	-494.7	-483.9	-508.4
ICL	-507.9	-469.7	-506.3	-492.6	-523.5

Valeurs du BIC et de l'ICL pour les différents nombres de groupes



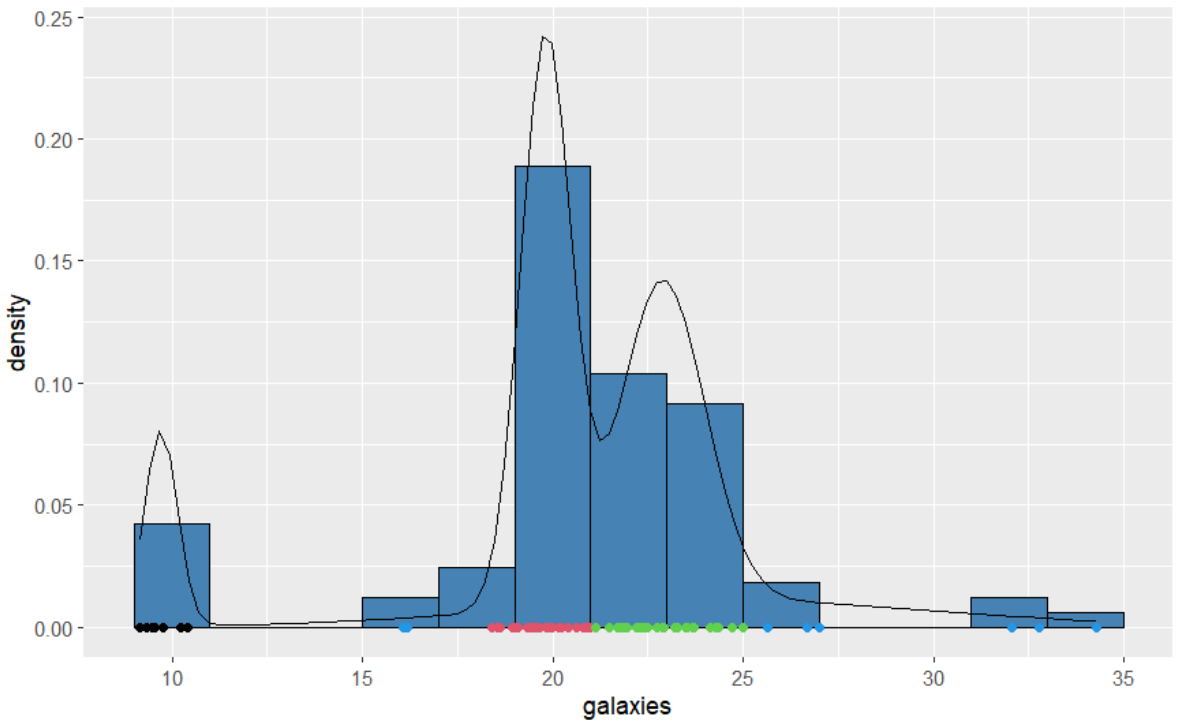
Visualisation finale avec les groupes d'appartenance et la densité de la mixture gaussienne pour K = 3

Plan

Comparaisons : Algorithme EM

Une autre méthode célèbre qui utilise les mixtures gausiennes pour faire du clustering est la méthode EM.
Comparons nos résultats avec celle-ci :

$$\hat{\pi} = (0.0844, 0.386, 0.370, 0.16)$$
$$\hat{\mu} = (9.71, 19.804, 22.878, 24.435)$$
$$\hat{\sigma^2} = (0.177, 0.435, 1.253, 34.122)$$



Visualisation du clustering par l'algorithme EM

Conclusion

Notre approche nous a permis de choisir un nombre à 3. Cela correspond à la littérature scientifique. Cela montre la viabilité de la méthode dans ce cas, cependant il faudrait la soumettre à des épreuves plus importantes avec plus de données et avec un nombre de dimensions plus élevé.

Merci de votre écoute !

Bibliographie I

Aitkin, M. (2001). Likelihood and bayesian analysis of mixtures. *Statistical Modelling*, 1(4) :287–304.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science : with applications in R*, volume 50. Cambridge University Press.

Conn, P. B., Johnson, D. S., Williams, P. J., Melin, S. R., and Hooten, M. B. (2018). A guide to bayesian model checking for ecologists. *Ecological Monographs*, 88(4) :526–542.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Grün, B., Malsiner-Walli, G., and Frühwirth-Schnatter, S. (2022). How many data clusters are in the galaxy data set ? bayesian cluster analysis in action. *Advances in data analysis and classification*, 16(2) :325–349.

Bibliographie II

Robert, C. P. et al. (2007). *The Bayesian choice : from decision-theoretic foundations to computational implementation*, volume 2. Springer.