Statistiques bayesiennes - Projet

M2 SSD

Benjamin Fayolle

UGA - Année universitaire 2019/2020

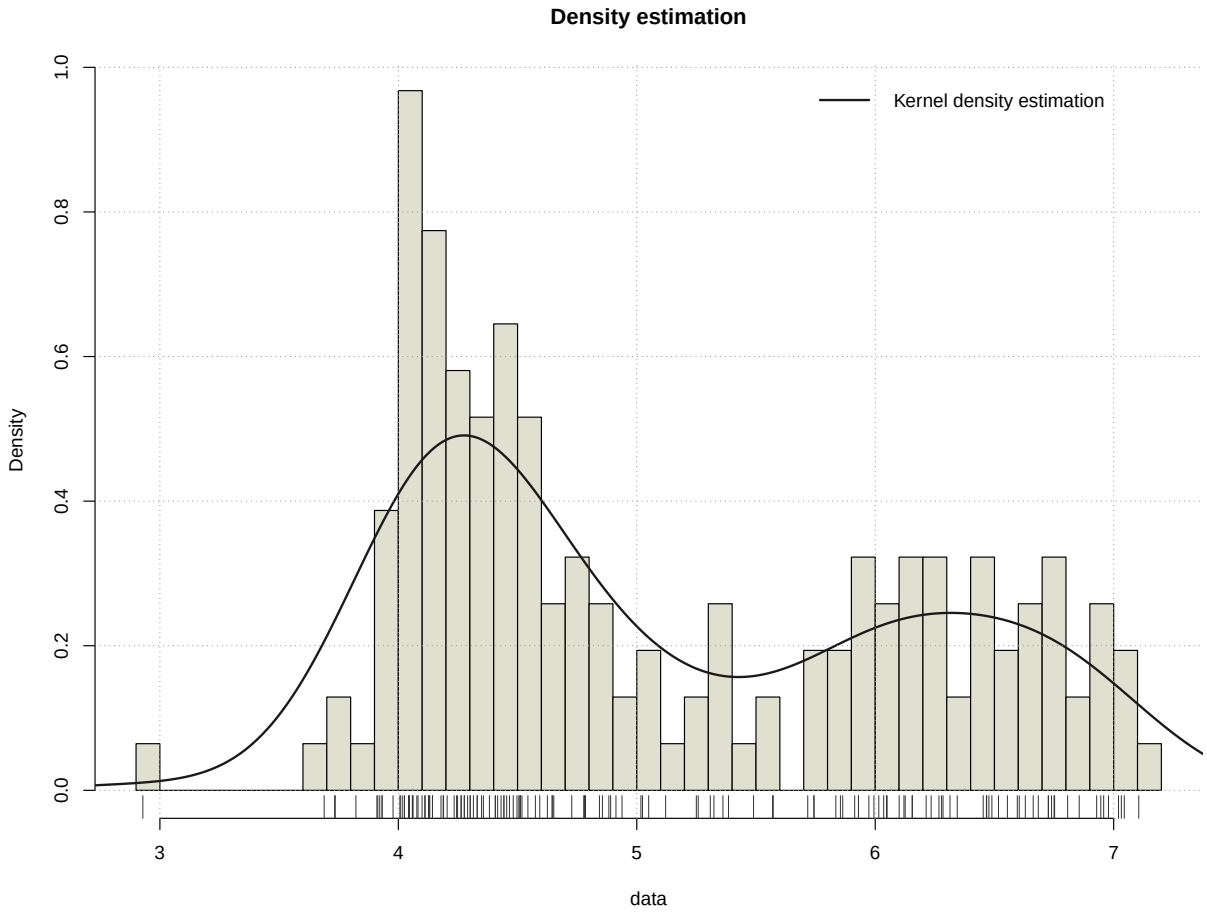


Figure 1: **Histogramme et densité estimée des donnée**. L'estimation de densité est faite via un noyeau gaussien.

Introduction

L'objectif de ce projet est l'estimation de paramètres en utilisant des méthodes bayesiennes, soit non seulement l'estimation de valeurs de paramètres à partir d'un jeu de données, mais aussi l'établissement d'une distribution de probabilité a posteriori pour ces derniers. Pour cela, nous nous reposerons sur des données mesurant un indice d'acidité dans un échantillon de 155 lacs de l'Amérique du Nord. Notons tout de suite que cet indice nous est donné en log, et sera traité comme tel. Nous sommes donc face à des données uni-dimensionnelles, prenant des valeurs comprises entre 2.9 et 7.1. La Figure 1 montre un histogramme des données ainsi qu'une estimation par noyeau de la densité.

On remarque, à la fois sur l'histogramme et sur la densité estimée, deux "bosses" que nous sommes tenté d'associer à deux gaussiennes différentes. Cette tentation est d'autant plus grande que cette approche a été adoptée par Crawford et all. [1] dans leurs analyses de 1992. Nous ferons ainsi l'hypothèse que la

distribution de nos données peut être modélisée par un mélange de deux gaussiennes, de moyennes, et semble-t-il d'écart-types, différents. Nous formulerons donc ainsi une première hypothèse concernant la distribution de nos données, et plus précisemment leur densité f :

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x), \tag{1}$$

où f_1 et f_2 désignent respectivement les densités de deux gaussiennes $\mathcal{N}(\mu_1, \sigma_1^2)$ et $\mathcal{N}(\mu_2, \sigma_2^2)$, et π_1 et π_2 sont les poids respectifs de ces deux gaussiennes.

Ceci étant posé, il est de notoriété publique qu'une manière efficace et largement utilisée afin d'estimer les paramètres d'un modèle de mélange, notamment de gaussienne, est l'algorithme *Espérance - Maximisation*, abrégé algorithme **EM**. Ce dernier permet d'estimer de manière relativement simple les paramètres permettant de maximiser la log-vraisemblance du modèle de mélange décrit plus haut, donnée (dans notre cas) par :

$$L(x, \Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{2} \pi_k f(x_i, \mu_k, \sigma_k) \right) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{2} \pi_k f_k(x_i) \right), \tag{2}$$

où L correspond à la log-vraisemblance, et $\Theta = (\pi_1, \ldots, \sigma_2)$ et le vecteur des paramètres du modèle. La maximisation directe de cette fonction étant trop complexe, l'algorithme **EM** procède en introduisant une variable indicatrice z_{ik} valant 1 si l'observation x_i appartient au groupe k, 0 sinon. Sans rentrer tout de suite dans les détail de l'algorithme, celui-ci a néanmoins l'inconvénient de recourir à une approche fréquentiste pour l'estimation des paramètres, ce qui ne nous convient bien sûr pas réellement.

Notre approche consistera ainsi à modifier l'algorithme **EM** afin d'inclure une étape d'estimation bayesienne des paramètres, tout en gardant les avantages de celui-ci.

Présentation de la méthode

L'algorithme que nous mettrons en œuvre afin d'analyser ces données peut être représenter shématiquement par l'algorithme 1. Nous rentrerons dans les détails de chaque étape plus loin.

Algorithme 1 Modification de l'algorithme EM

procédure EM_BAYES(données)
    Initialisation ▷ (a)
    tant que les paramètres n'ont pas convergé **faire** ▷ (b)
        1. Réaffecter les observations dans les deux groupes
        2. Estimer les paramètres des deux groupes via un algorithme de Gibbs
        3. Estimer les paramètres optimaux via les formules classiques de l'algorithme **EM**, qui serviront à initialiser le Gibbs de la prochaine tération
        4. Actualiser le critère de convergence des paramètres
    fin tant que
    Sortie Les résultats du Gibbs de la dernière itération
fin procédure

Etape d'initialisation (a)

Nous avons d'ors et déjà fait l'hypothèse que les données étaient déparées en deux groupes. Nous commencerons donc par affecter chaque observation à un groupe. Lors de cette étape d'initialisation, l'affectation sera réalisée au hasard, avec équiprobabilité pour chaque groupe, et donc $\forall i \in [\![1, \dots, n]\!]$, $P(x_i \in G_1) = P(x_i \in G_2) = 0.5$. Ainsi, lors de l'initialisation, $\pi_1 = \pi_2 = 0.5$. Une fois les deux groupes construits, il nous faut en estimer les paramètres. Rappelons que nous avons fait l'hypothèse que chaque groupe était distribué selon une gaussienne de moyenne et de variance inconnues. Nous sommes donc dans un cas d'estimation multi-dimensionnel. Nous sommes de plus dans le cas gaussien, qui est un cas classique. Nous auront donc recours aux lois a priori classiques de ce cas. Pour résumer :

- $G_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$, où $\theta_k = (\mu_k, \sigma_k^2)$ est inconnu
- La loi a priori de μ_k est une gaussienne $\mathcal{N}(\mu_{0k}, \sigma_{0k}^2)$
- La loi a priori de σ_k^2 est une Inverse Gamma $IG(a_k, b_k)$

Nous estimons ainsi θ_k grâce à un algorithme de Gibbs [2], en prenant comme paramètres initiaux les estimateurs empiriques de la moyenne et de la variance sur chacun des groupes. Pour la loi Inverse Gamma, nous calculons les paramètres a_k et b_k grâce à leur relation avec la moyenne et la variance, que nous noterons ν et τ^2 :

$$a_k = \frac{\nu_k^2}{\tau_k^2} + 2; \quad b_k = \nu_k(a_k - 1)$$

Les paramètres de variances de nos deux prior sont fixé arbitrairement à 1. Cela étant, une variation de ce paramètre n'entraine quasiment aucune variation de l'estimation en sortie de l'algorithme complet. Pour résumer, les paramètres de départ des deux algorithmes Gibbs de la phase d'initialisation sont :

- $\mu_k^{(0)} = \bar{X}_k$
- $\sigma_k^{2(0)} = 1$
- $a_k^{(0)} = S_n^2 + 2$
- $b_k^{(0)} = S_n(a_k^{(0)} - 1)$

avec $\bar{X}_k$ la moyenne empirique de chacun des deux groupes, et S_n leur variances empiriques. A l'issue des algorithmes de Gibbs nous obtenons les lois a posteriori de μ_k et σ_k^2. Nous prenons ainsi comme estimation de ces derniers la moyenne de leur distribution respectives. Vient ensuite l'étape d'optimisation des paramètres de l'algorithme **EM**. Cette étape consiste à trouver la combinaison de paramètres qui maximise la quantité

$$Q = \sum_{i=1}^{n} \sum_{k=1}^{2} t_{ik} \log\left(\pi_k f_k(x_i)\right), \tag{3}$$

où

$$t_{ik} = \frac{\pi_k f_k(x_i)}{\sum_{\ell=1}^{2} \pi_\ell f_\ell(x_i)} \tag{4}$$

Comme nous sommes dans le cas gaussien, les paramètres optimaux sont donnés par :

$$\pi_k^* = \frac{1}{n}\sum_{i=1}^{n} t_{ik}; \quad \mu_k^* = \frac{\sum_{i=1}^{n} t_{ik} x_i}{\sum_{i=1}^{n} t_{ik}}; \quad \sigma_k^{2*} = \frac{\sum_{i=1}^{n} t_{ik}(x_i - \mu_k^*)^2}{\sum_{i=1}^{n} t_{ik}} \tag{5}$$

Bien entendu, cette étape d'estimation est purement fréquentiste, en ce qu'elle consiste en réalité à maximiser l'espérance de la log-vraisemblance du modèle. Toutefois, les paramètres ainsi estimés serviront uniquement à initialiser les algorithmes Gibbs des prochaines itérations. Ainsi, seules les proportions π_k^* seront directement utilisées. La phase d'initialisation de notre méthode est donc terminée.

Boucle (b)

Pour la majeur partie, les itérations de la méthode proposée reprennent les étapes de l'initialisation, du moins conceptuellement. La seule réelle nouveauté est la réaffectation de chaque individu dans un groupe. En

effet, les groupes initiaux étant aléatoires, il nous faut petit à petit tendre vers les "vrais" groupes. Pour cela, au début de chaque itération, on reconstruit les groupes grâce aux t_{ik}. En effet, ces derniers peuvent être interprétés comme la probabilité, pour une observation, d'appartenir au groupe k. Ainsi, chaque observation x_i est affectée au groupe 1 si $t_{i1} > 0.5$, et au groupe 2 sinon. Une fois que nous avons les groupes, nous devons à nouveau estimer les paramètres (μ_k, σ_k^2) en utilisant un algorithme de Gibbs. La seule différence avec l'étape d'initialisation est que les paramètres de départ de l'algorithme Gibbs ne sont plus calculés sur l'échantillon, mais correspondent à $(\mu_k^*, \sigma_k^{2^*})$ calculés précédemment. Les proportions π_k sont elles remplacées par les proportions optimales π_k^*.

Nous utilisons ensuite les paramètres estimés par Gibbs pour actualiser les t_{ik}, puis nous actualisons les paramètres optimaux $(\pi_k^*, \mu_k^*, \sigma_k^{2^*})$, ainsi que la qauntité Q. Enfin, nous mettons à jours le critère de convergence, avant de commencer, si besoin, une nouvelle itération.

Concernant ce critère de convergence, noté C, nous l'avons défini comme étant la différence (en valeur absolue) entre l'ancienne quantité Q, définie précédemment, et la nouvelle :

$$C = |Q^{new} - Q^{old}| \tag{6}$$

L'idée est simple : comme Q représente l'espérance de la log-vraisemblance du modèle, alors Q devrait prendre des valeurs de plus en plus grandes à mesure que le modèle converge. Au bout d'un moment, l'augmentation de Q ne sera qu'extrememement marginale, puisque les paramètres trouvés ne pourront guère être améliorés. Dès lors, nous définissons le critère d'arrêt suivant :

$$C < \varepsilon, \tag{7}$$

où ε est un seuil choisit empiriquement. Notons que, dans le cas présent, le modèle converge très rapidement. La Figure 2 montre en ce sens l'évolution de la quantité Q au fil des itérations du modèle. Pour ce projet, nous avons fixé ε de manière empirique à 0.001.

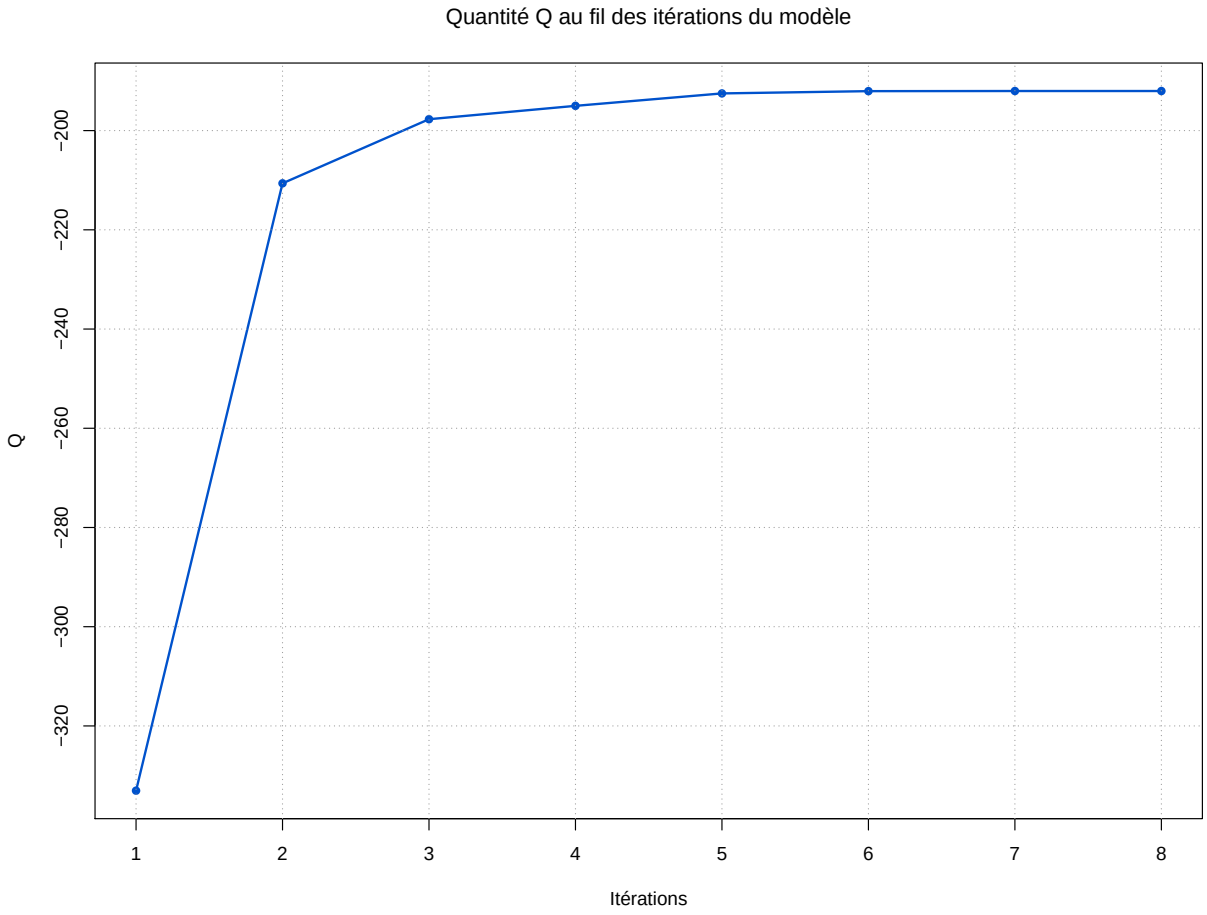
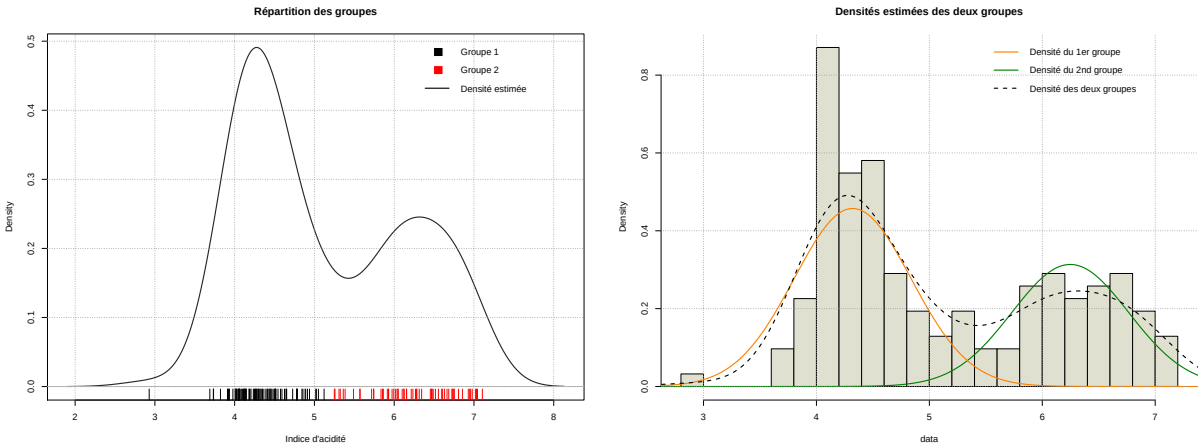


Figure 2: **Vitesse de convergence du modèle**.

Résultats

A l'issue de cette méthode, nous obtenons les estimations de tous les paramètres que nous cherchions. Ainsi, μ_1, μ_2, σ_1, σ_2 sont estimés par l'algorithme de Gibbs lors de la dernière itération de la méthode avant convergence. Les proportions π_1 et π_2 sont, elles, estimées de manière fréquentiste par l'algorithme **EM**. La méthode permet au passage de réaliser un clustering non supervisé, dans la mesure où elle affecte chaque observation dans l'un des deux groupes. Ainsi, les graphiques de la Figure 3 montrent respectivement la répartition des groupes ainsi que la densité estimée des données en fonction de la valeur de l'indice d'acidité (a) et les densitées estimées des deux groupes obtenues comparées à la densité commune (b).



(a) Répartition des groupes

(b) Densités des deux groupes

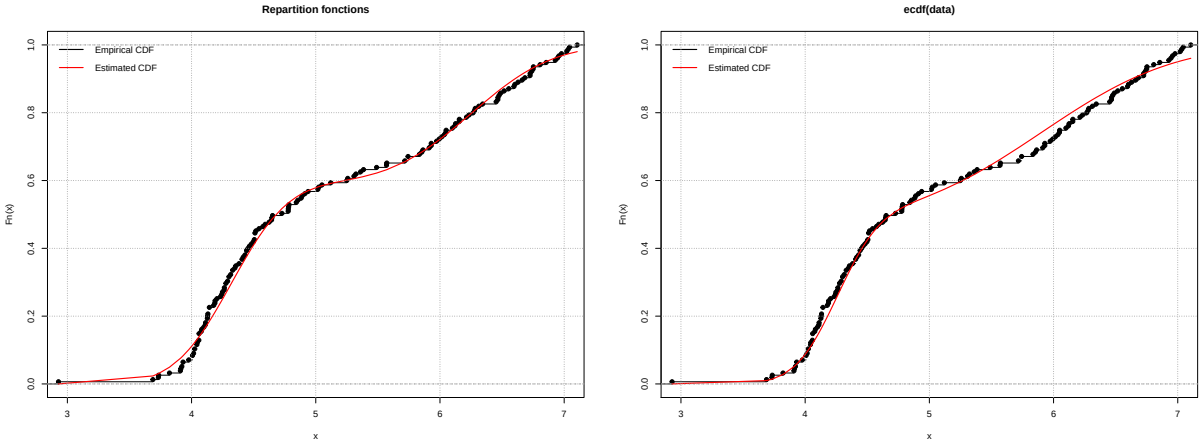
Figure 3: **Représentation des groupes obtenus après convergence de la méthode**. Dans le second graphique, chaque densité à été multipliée par la proportion π_k associée afin d'être à la bonne échelle.

La table 1 montre quant à elle les estimations réalisées de chacun des paramètres du modèle. Elle porpose aussi la valeur des estimations réalisées par un algorithme **EM** simple, à des fins de comparaison.

Table 1: Estimations des paramètres

	Méthode proposée		Algo EM simple	
	Groupe 1	Groupe 2	Groupe 1	Groupe 2
μ	4.32042	6.247814	4.25291	5.901305
σ	0.3600304	0.517549	0.8420562	0.2626358
π	0.5932842	0.4067158	0.4830206	0.5169794

Enfin, la Figure 4 donne une idée de l'adéquation du modèle aux données, en comparant la fonction de répartition empirique des données à celle du modèle estimé (a). Le second graphique donne la même représentation pour un algorithme **EM** simple, a des fins de comparaison.



(a) Méthode proposée

(b) Algorithme EM simple

Figure 4: **Comparaison des fonctions de répartition empirique et estimée**.

Références

[1] Sybil L. Crawford, Morris H. DeGroot, Joseph B. Kadane, and Mitchell J. Small. Modeling lake-chemistry distributions: Approximate bayesian methods for estimating a finite-mixture model. *Technometrics*, 34(4):441–453, 1992. http://www.jstor.org/stable/1268943.

[2] B Walsh. Markov chain monte carlo and gibbs sampling. *Lecture Notes for EEB 581, version 26, April*, 01 2004.