



Projet Master 1 SSD

Un modèle pour les nids d'oiseaux

Rédigé par

CARVAILLO Thomas

CÔME Olivier

PRALON Nicolas

Encadrante : Elodie BRUNEL-PICCININI

IMAG
INSTITUT MONTPELLIERAIN
ALEXANDER GROTHENDIECK



Table des matières

Introduction	2
1 Modélisation du problème et liminaire mathématique	3
1.1 Modélisation du problème	3
1.2 Une histoire de densités	4
1.3 Une approche idéaliste	5
1.4 Une situation concordante à la réalité	7
2 L'algorithme EM	12
2.1 Présentation laconique et pseudo-code	12
2.1.1 L'étape E (Expectation)	12
2.1.2 L'étape M (Maximization)	12
2.1.3 La fonction simulation	12
2.1.4 Pseudo-code de l'algorithme EM	13
2.2 Une preuve de la croissance	14
3 Une mise en situation	16
3.1 Préambule	16
3.2 Étude de simulations	17
3.2.1 Étude d'un mélange gaussien à forte séparations	17
3.2.2 Étude d'un mélange gaussien à faible séparations	19
3.3 Étude de cas avec les vrais données	22
3.3.1 Cas d'un mélange à deux lois avec des vrais données	22
3.4 Cas d'un mélange à trois lois avec des vrais données	24
Bibliographie	25
Annexes	26
A Le package <i>mclust</i>	27
A.1 Un exemple sur un mélange à deux lois	28
A.2 Un exemple sur un mélange à trois lois	30

Introduction

L'observation est une composante essentielle en sciences appliquées, et tout particulièrement pour nous, futur statisticiens. Ces dernières nous permettent d'élaborer des modèles, et dans le cadre de l'estimation paramétrique, de construire des estimateurs.

Dans la situation où les données sont correctement observées, cadre idéal, la statistique inférentielle nous permet d'obtenir d'agréables expressions analytiques des estimateurs, notamment par la méthode du maximum de vraisemblance, qui est d'une redoutable efficacité. Toujours est-il que ces situations sont peu courantes; nous sommes souvent confrontés à des situations présentant des données « cachées » ou manquantes. Dans ce dernier cas, nous verrons que les méthodes analytiques exactes ne suffisent plus, et qu'il est nécessaire d'introduire de nouvelles méthodes, basées sur le principe du maximum de vraisemblance, pour tenter d'estimer les paramètres.

Nous étudierons dans le présent projet l'une de ces méthodes, conçue par Dempster et al. (1977), l'algorithme Expectation-Maximization.

Nous nous appuyerons ici sur un exemple extrait de l'ornithologie; les données observées seront celles de la taille des nids d'oiseaux, et les données manquantes seront l'espèce qui l'a construit.

Dans une première partie, nous modéliserons mathématiquement le problème et poserons les bases théoriques nécessaires à l'étude; nous verrons brièvement le cas des données complètes, puis étudierons le cas des observations manquantes. Une seconde partie sera consacrée à une implémentation en langage R de l'algorithme EM et aux performances des résultats de ce dernier. Nous détaillerons notre implémentation, et étudierons son risque d'erreur. Dans une troisième et dernière partie, nous nous proposons de nous mettre en conditions réelles et de réaliser une étude de cas à l'aide des outils que nous aurons implémentés.

Chapitre 1

Modélisation du problème et liminaire mathématique

Dans ce premier chapitre, nous introduirons les outils nécessaires pour modéliser mathématiquement le problème. Puis, nous présenterons successivement les deux situations existantes dans le cadre du recueil de données des nids d'oiseaux : celle dans laquelle et l'espèce et la taille du nid ont été observées ; et celle dans laquelle seule la taille a pu être observée. Nous verrons et ce qui diffère entre ces deux situations, et les raisons qui ont incité à la création de l'algorithme EM, l'objet de notre projet.

Ce chapitre s'inspirera indifféremment des références [1], [2] et [3]

1.1 Modélisation du problème

Nous allons pour commencer donner une première définition, qui est au coeur du présent projet.

Définition 1 (Loi de mélange). On appelle loi de mélange toute loi dont la densité s'écrit sous la forme d'une combinaison convexe de plusieurs densités. Si l'on se donne J densités $f_1(x), \dots, f_J(x)$, alors toute variable aléatoire X dont la densité f s'exprime, pour tout $x \in \mathbb{R}$, sous la forme

$$f(x) := \sum_{i=1}^J \alpha_i f_i(x), \alpha_i \in \mathbb{R}_+^* \text{ et } \sum_{j=1}^J \alpha_j = 1$$

suit une loi de mélange.

Afin de modéliser commodément le problème, nous introduisons les variables aléatoires suivantes :

- ✂ La variable aléatoire X , modélisant le volume des nids, de densité f
- ✂ Z , la variable aléatoire discrète et à valeurs dans $\llbracket 1, J \rrbracket$, représentant l'espèce d'oiseau qui a construit le nid

Enfin, nous nous placerons sous les hypothèses suivantes :

Hypothèse 1. Nous supposons que, $\forall j \in \llbracket 1, J \rrbracket$, la taille des nids d'une espèce j (i.e. X conditionnellement à $(Z = j)$) suit une loi normale $\mathcal{N}(\mu_j, v_j)$. Nous dénoterons par $f(x|Z = j) := \gamma_{\mu_j, v_j}(x)$ cette densité.

Hypothèse 2. Soit $\Theta := \{\theta = (\alpha_j, \mu_j, v_j)_{1 \leq j \leq J} \text{ tels que } \alpha_j > 0 \forall j \in \llbracket 1, J \rrbracket \text{ et } \sum_{j=1}^J \alpha_j = 1\}$. Soient X_1, \dots, X_n un échantillon de même loi que X . On supposera qu'il existe un $\theta \in \Theta$ tel que les données récoltées, ici les tailles des nids, soient la réalisation du précédent échantillon.

Remarque 1. La variable Z est discrète et à valeur dans un sous-ensemble fini de \mathbb{N} , elle suit donc une loi

$$\sum_{j=1}^J \alpha_j \delta_j$$

où J représente le nombre d'espèce de d'oiseaux considéré et les α_j sont des réels, positifs stricts, représentant la proportion de nids de l'espèce j , tels que $\sum_{j=1}^J \alpha_j = 1$.

Il s'ensuit la proposition suivante, qui sera essentielle dans la suite.

Proposition 1. *La distribution de la taille des nids des oiseaux, i.e. X , admet pour densité, au point x et par rapport à la mesure de Lebesgue sur \mathbb{R} , la fonction f_θ définie comme suit*

$$f_\theta(x) = \sum_{j=1}^J \alpha_j \gamma_{\mu_j, v_j}(x)$$

Démonstration. En effet,

$$\begin{aligned} \mathbb{P}(X \leq x) &= \mathbb{P}\left(\bigcup_{j=1}^J (X \leq x) \cap (Z = j)\right) \\ &= \sum_{j=1}^J \mathbb{P}((X \leq x) \cap (Z = j)) \\ &= \sum_{j=1}^J \mathbb{P}(Z = j) \times \mathbb{P}(x \leq X | Z = j) \\ &= \sum_{j=1}^J \alpha_j \times \mathbb{P}(x \leq X | Z = j) \\ &= \sum_{j=1}^J \alpha_j \times f(x | Z = j) \end{aligned}$$

2

Le but de ce projet sera d'étudier des méthodes permettant l'estimation des divers paramètres de cette densité. Nous dénoterons par $\theta := (\alpha_j, \mu_j, v_j)_{1 \leq j \leq J}$ le vecteur des paramètres.

1.2 Une histoire de densités

Introduisons une dernière densité et une dernière probabilité, qui nous seront fort utile dans la suite :

Proposition 2. *Nous avons les résultats suivant :*

1. *La loi du couple (X, Z) est donnée par*

$$\begin{aligned} \mathbb{P}(X \leq x, Z = j) &= \int_{-\infty}^x f_\theta(u | Z = j) \times \mathbb{P}(Z = j) du \\ &= \int_{-\infty}^x \underbrace{\gamma_{\mu_j, v_j}(u) \times \alpha_j}_{:= h_\theta(u, j)} du \end{aligned}$$

où $h_\theta(u, j)$ est la "sous-densité" de $X \times \mathbb{1}_{(Z=j)}$

2. *La probabilité de la loi de Z sachant $X = x$ est donnée par :*

$$\mathbb{P}_\theta(Z = j | X = x) = \frac{\gamma_{\mu_j, v_j} \times \alpha_j}{f_\theta(x)}$$

où $f_\theta(x)$ est donnée par la proposition 1.

Démonstration. En effet,

$$\mathbb{P}(X \leq x, Z = j) = \int_{-\infty}^x \mathbb{P}(Z = j | X = u) \times f_\theta(u) du$$

On obtient dès lors

$$\mathbb{P}(Z = j|X = u) \times f_\theta(u) = \gamma_{\mu_j, v_j}(u) \times \alpha_j$$

Soit

$$\mathbb{P}(Z = j|X = u) = \frac{\gamma_{\mu_j, v_j}(u) \times \alpha_j}{f_\theta(u)} := \frac{h_\theta(u, j)}{f_\theta(u)}$$

2

Remarque 2. Nous pouvons dès à présent noter que pour un échantillon X_1, \dots, X_n de même loi que X , nous avons

$$\forall i \in \llbracket 1, n \rrbracket, h_\theta(X_i, j) = f_\theta(X_i) \times \mathbb{P}(Z = j|X = X_i)$$

Ceci nous sera utile dans la suite.

Nous allons dès à présent nous intéresser à l'estimation de ces paramètres.

1.3 Une approche idéaliste

Regardons dans un premier temps un cas simplifié, un cas ne décrivant pas la réalité des observations mais qui constitue une agréable entrée en matière.

Nous supposons ici qu'ont été relevés simultanément et les mesures des tailles des nids et l'espèce d'oiseau qui l'a construit. Les observations considérées ici sont donc composées des couples (X_i, Z_i) , $i \in \llbracket 1, n \rrbracket$. On considérera dès lors la fonction de densité $h_\theta(x, z)$, donnée par la proposition 2.

L'estimation des divers paramètres est alors élémentaire, en témoigne les propositions suivantes :

Proposition 3 (Fonction de Log-vraisemblance). *La Log-vraisemblance du modèle s'écrit*

$$\mathcal{L}_\theta(X_1, \dots, X_n, Z_1, \dots, Z_n) = \sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln(\gamma_{\mu_j, v_j}(X_i))$$

où les A_j sont définis par $A_j := \{i \in \llbracket 1, n \rrbracket \text{ tels que } Z_i = j\}$ i.e. $\bigcup_{j=1}^J A_j = \llbracket 1, n \rrbracket$

Avant de démontrer cette proposition, nous allons introduire une notation qui nous sera immédiatement utile :

Notation 1. Dans ce qui suit, nous noterons

$$\delta_j := \mathbb{1}_{(Z_i=j)}(Z_i)$$

Ainsi,

$$h_\theta(X_i, Z_i) = \prod_{j=1}^J h_\theta(X_i, Z_i)^{\delta_j}$$

Démonstration. La Log-vraisemblance du modèle s'écrit :

$$\begin{aligned} \mathcal{L}_\theta(X_1, \dots, X_n, Z_1, \dots, Z_n) &= \ln \left(\prod_{i=1}^n h_\theta(X_i, Z_i) \right) \\ &= \ln \left(\prod_{i=1}^n \prod_{j=1}^J h_\theta(X_i, Z_i)^{\delta_j} \right) \\ &= \mathbb{1}_{(Z_i=j)}(Z_i) \times \ln \left(\prod_{i=1}^n \prod_{j=1}^J h_\theta(X_i, Z_i) \right) \end{aligned}$$

Z_i est à valeur dans $j \in \llbracket 1, J \rrbracket$, on partitionne donc $I := \llbracket 1, n \rrbracket$ comme $I = \bigcup_{j=1}^J A_j$.

Ceci va nous permettre de nous désencombrer de l'indicatrice en réindexant la somme. Nous obtenons dès lors :

$$\begin{aligned}
\mathcal{L}_\theta(X_1, \dots, X_n, Z_1, \dots, Z_n) &= \ln \left(\prod_{i \in A_i} \prod_{j=1}^J h_\theta(X_i, Z_i) \right) \\
&= \ln \left(\prod_{i \in A_i} \prod_{j=1}^J \alpha_j \gamma_{\mu_j, v_j}(X_i) \right) \\
&= \sum_{i \in A_i} \sum_{j=1}^J \ln(\alpha_j \gamma_{\mu_j, v_j}(X_i)) \\
&= \sum_{j=1}^J \sum_{i \in A_j} \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln(\gamma_{\mu_j, v_j}(X_i)) \\
&= \sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln(\gamma_{\mu_j, v_j}(X_i))
\end{aligned}$$

2

Nous pouvons dès lors maximiser la log-vraisemblance afin d'obtenir les estimateurs souhaités :

Proposition 4 (Estimateurs). *Les estimateurs du maximum de vraisemblance $\widehat{\alpha}_j$ (resp. $\widehat{\mu}_j$, et \widehat{v}_j) de α_j (resp. μ_j et v_j) sont donnés par*

$$\begin{aligned}
\widehat{\alpha}_j &= \frac{\#A_j}{n} \\
\widehat{\mu}_j &= \frac{\sum_{i \in A_j} X_i}{\#A_j} \\
\widehat{v}_j &= \frac{\sum_{i \in A_j} (X_i - \widehat{\mu}_j)^2}{\#A_j}
\end{aligned}$$

Démonstration. Soit $\theta = (\alpha_j, \mu_j, v_j)_{j \in \llbracket 1, J \rrbracket}$. Il s'agit de déterminer

$$\operatorname{argmax}_{\theta \in \mathbb{R}^{3J}, \sum_{j=1}^J \alpha_j = 1} \left(\sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln(\gamma_{\mu_j, v_j}(X_i)) \right)$$

Nous avons donc à résoudre un programme de minimisation d'une fonction convexe sur un convexe avec une contrainte égalité, il est ainsi naturel de faire appel au Lagrangien.

Ce dernier s'écrit

$$\begin{aligned}
L(\theta) &= \sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln(\gamma_{\mu_j, v_j}(X_i)) - \lambda \times \left(\sum_{j=1}^J \alpha_j - 1 \right) \\
&= \sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \ln \left(\frac{1}{\sqrt{2\pi v_j}} \exp \left(-\frac{(X_i - \mu_j)^2}{2v_j} \right) \right) - \lambda \times \left(\sum_{j=1}^J \alpha_j - 1 \right) \\
&= \sum_{j=1}^J \#A_j \ln(\alpha_j) + \sum_{j=1}^J \sum_{i \in A_j} \left(\frac{-1}{2} \ln(2\pi v_j) - \frac{(X_i - \mu_j)^2}{2v_j} \right) - \lambda \times \left(\sum_{j=1}^J \alpha_j - 1 \right)
\end{aligned}$$

Il reste maintenant à résoudre le système suivant, afin d'obtenir le vecteur $\widehat{\theta} := (\widehat{\alpha}_j, \widehat{\mu}_j, \widehat{v}_j)_{j \in \llbracket 1, J \rrbracket}$ solution du programme.

$$\begin{cases} \frac{\#A_j}{\widehat{\alpha_j}} - \lambda & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} (X_i - \widehat{\mu_j}) / \widehat{v_j} & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} \frac{-0.5 \times 2 \times \pi}{2\pi \widehat{v_j}} + \frac{(X_i - \widehat{\mu_j})^2}{2\widehat{v_j}^2} & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{j=1}^J \widehat{\alpha_j} & = 1 \end{cases}$$

Ceci équivaut à

$$\begin{cases} \frac{\#A_j}{\widehat{\alpha_j}} & = \lambda \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} X_i & = \sum_{i \in A_j} \widehat{\mu_j} \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} (X_i - \widehat{\mu_j})^2 & = \sum_{i \in A_j} \widehat{v_j} \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{j=1}^J \widehat{\alpha_j} & = 1 \end{cases} \Leftrightarrow \begin{cases} \frac{\#A_j}{\widehat{\alpha_j}} & = \lambda \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} \frac{X_i}{\#A_j} & = \widehat{\mu_j} \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i \in A_j} \frac{(X_i - \widehat{\mu_j})^2}{\#A_j} & = \widehat{v_j} \quad \forall j \in \llbracket 1, J \rrbracket \\ \sum_{j=1}^J \widehat{\alpha_j} & = 1 \end{cases}$$

En sommant les J premières lignes du système, on obtient $\sum_{j=1}^J \#A_j = \sum_{j=1}^J \widehat{\alpha_j} \lambda$, i.e. $\lambda = n$. En injectant ceci dans le précédent système, on obtient finalement ce qui était annoncé :

$$\begin{cases} \widehat{\alpha_j} & = \frac{\#A_j}{n} \quad \forall j \in \llbracket 1, J \rrbracket \\ \widehat{\mu_j} & = \sum_{i \in A_j} \frac{X_i}{\#A_j} \quad \forall j \in \llbracket 1, J \rrbracket \\ \widehat{v_j} & = \sum_{i \in A_j} \frac{(X_i - \widehat{\mu_j})^2}{\#A_j} \quad \forall j \in \llbracket 1, J \rrbracket \end{cases}$$

1.4 Une situation concordante à la réalité

Nous nous placerons désormais dans un contexte tout autre que celui du paragraphe précédent, un contexte concordant davantage à la réalité. Dans ce qui suit, nous supposons que ne sont observées que les tailles des nids, les diverses espèces d'oiseaux les ayant construit étant en quelque sorte des données inobservées ou "cachées". Nous avons donc un échantillon X_1, \dots, X_n de même loi que la variable X comme définie ci-dessus. La log-vraisemblance des observations \mathcal{L}_{obs} s'obtient aisément :

$$\mathcal{L}_{obs}(\theta, X_1, \dots, X_n) := \ln \left(\prod_{i=1}^n f_{\theta}(X_i) \right) = \sum_{i=1}^n \ln \left(\sum_{j=1}^J \alpha_j \gamma_{\mu_j, v_j}(X_i) \right)$$

Nous voyons dès lors que l'existence d'une expression analytique du maximum de la log-vraisemblance n'est pas assurée. Il est donc nécessaire de trouver un moyen d'approcher les valeurs des différents estimateurs.

Pour ce faire, on définit une log-vraisemblance des couples (X_i, Z_i) sachant le vecteurs des observations X_1, \dots, X_n .

Définition 2 (log-vraisemblance conditionnelle). On définit la log-vraisemblance $\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n)$ conditionnelle par

$$\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) = \mathbb{E}_{\tilde{\theta}}[\mathcal{L}_{\theta}(X_1, \dots, X_n, Z_1, \dots, Z_n) | X_1, \dots, X_n]$$

Nous allons maintenant travailler sur l'expression de la log-vraisemblance conditionnelle et en donner une expression simplifiée, qui nous sera fort utile ultérieurement, et une expression plus substantielle, qui nous sera immédiatement utile.

Proposition 5. *Nous avons*

$$\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) = \sum_{i=1}^n \sum_{j=1}^J \ln(h_\theta(X_i, j)) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)$$

Démonstration. En effet

$$\begin{aligned} \mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) &= \mathbb{E}_{\tilde{\theta}}[\mathcal{L}_\theta(X_1, \dots, X_n, Z_1, \dots, Z_n) | X_1, \dots, X_n] \\ &= \mathbb{E}_{\tilde{\theta}} \left[\ln \left(\prod_{i=1}^n \prod_{j=1}^J h_\theta(X_i, Z_i)^{\delta_j} \right) \middle| X_1, \dots, X_n \right] \\ &= \sum_{i=1}^n \sum_{j=1}^J \mathbb{E}_{\tilde{\theta}} [\delta_j \times \ln(h_\theta(X_i, Z_i)) | X_1, \dots, X_n] \\ &= \sum_{i=1}^n \sum_{j=1}^J \mathbb{E}_{\tilde{\theta}} [\mathbb{1}_{(Z_i=j)}(Z_i) \times \ln(h_\theta(X_i, Z_i)) | X_1, \dots, X_n] \end{aligned}$$

Or, les couples (X_i, Z_i) sont indépendant, donc

$$\begin{aligned} \mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) &= \sum_{i=1}^n \sum_{j=1}^J \mathbb{E}_{\tilde{\theta}} [\mathbb{1}_{(Z_i=j)}(Z_i) \times \ln(h_\theta(X_i, Z_i)) | X_i] \\ &= \sum_{i=1}^n \sum_{j=1}^J \mathbb{E}_{\tilde{\theta}} [\mathbb{1}_{(Z_i=j)}(Z_i) \times \ln(h_\theta(X_i, j)) | X_i] \\ &= \sum_{i=1}^n \sum_{j=1}^J \ln(h_\theta(X_i, j)) \mathbb{E}_{\tilde{\theta}} [\mathbb{1}_{(Z_i=j)}(Z_i) | X_i] \\ &= \sum_{i=1}^n \sum_{j=1}^J \ln(h_\theta(X_i, j)) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \end{aligned}$$

2

Nous nous appuyerons sur l'expression suivante pour l'expression des estimateurs du maximum de vraisemblance :

Proposition 6. La fonction $\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n)$ se réécrit sous la forme suivante :

$$\begin{aligned} \mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) &= -\frac{n}{2} \ln(2\pi) + \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \right) \ln(\alpha_j) \\ &\quad - \frac{1}{2} \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \left(\log(v_j) + \frac{(X_i - \mu_j)^2}{v_j} \right) \right) \end{aligned}$$

Démonstration. Il suffit de partir de la forme précédente de la log-vraisemblance conditionnelle, on a ainsi :

$$\begin{aligned}
\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) &= \sum_{i=1}^n \sum_{j=1}^J \ln(h_{\theta}(X_i, j)) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J \ln(\alpha_j \gamma_{\mu_j, v_j}(X_i)) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J (\ln(\alpha_j) + \ln(\gamma_{\mu_j, v_j}(X_i))) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J \ln(\alpha_j) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) + \sum_{i=1}^n \sum_{j=1}^J \ln(\gamma_{\mu_j, v_j}(X_i)) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)
\end{aligned}$$

Traitons pour commencer la double somme

$$\Delta := \sum_{i=1}^n \sum_{j=1}^J \ln(\gamma_{\mu_j, v_j}(X_i)) \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)$$

Nous avons :

$$\gamma_{\mu_j, v_j}(X_i) = \frac{1}{\sqrt{2\pi v_j}} e^{-\frac{1}{2} \frac{(X_i - \mu_j)^2}{v_j}}$$

et

$$\mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) = \frac{\alpha_j \gamma_{\mu_j, v_j}}{\sum_{k=1}^J \alpha_k \gamma_{\mu_k, v_k}}$$

La double somme devient alors

$$\begin{aligned}
\Delta &= \sum_{i=1}^n \sum_{j=1}^J \ln \left(\frac{1}{\sqrt{2\pi v_j}} e^{-\frac{1}{2} \frac{(X_i - \mu_j)^2}{v_j}} \right) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J \ln \left(\frac{1}{\sqrt{2\pi v_j}} \right) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) - \frac{1}{2} \left(\frac{(X_i - \mu_j)^2}{v_j} \right) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J -\frac{1}{2} \ln(2\pi) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) - \frac{1}{2} \ln(v_j) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) - \frac{1}{2} \left(\frac{(X_i - \mu_j)^2}{v_j} \right) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^J \left(\ln(v_j) + \frac{(X_i - \mu_j)^2}{v_j} \right) \times \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\
&= -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \times \left(\ln(v_j) + \frac{(X_i - \mu_j)^2}{v_j} \right) \right)
\end{aligned}$$

On obtient de fait le résultat espéré :

$$\begin{aligned}
\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n) &= -\frac{n}{2} \log(2\pi) + \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \right) \times \ln(\alpha_j) \\
&\quad - \frac{1}{2} \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \times \left(\ln(v_j) + \frac{(X_i - \mu_j)^2}{v_j} \right) \right)
\end{aligned}$$

Nous allons dès à présent énoncer une proposition vitale, celle de l'expression des estimateurs du maximum de vraisemblance de la log-vraisemblance conditionnelle. L'expression de ces derniers seront le pivot de l'algorithme EM, que nous présenterons dans le chapitre suivant.

Proposition 7. *La fonction $\theta \mapsto \mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n)$ admet un unique maximum θ_M donné par :*

$$\begin{aligned}\widehat{\alpha}_j &= \frac{1}{n} \sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \\ \widehat{\mu}_j &= \frac{\sum_{i=1}^n X_i \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}{\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)} \\ \widehat{v}_j &= \frac{\sum_{i=1}^n (X_i - \mu_j)^2 \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}{\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}\end{aligned}$$

Démonstration. Soit $\theta = (\alpha_j, \mu_j, v_j)$. Il s'agit ici de maximiser la fonction $\theta \mapsto \mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n)$

Puisqu'il s'agit d'un problème d'optimisation, nous appliquons la même méthode que précédemment, en introduisant le Lagrangien du problème sous la contrainte $\sum_{i=1}^n \alpha_i = 1$.

Nous reprenons ici l'écriture de $\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \dots, X_n)$ donnée dans la précédente proposition, nous obtenons ainsi l'expression suivante du Lagrangien

$$\begin{aligned}L(\theta, \lambda) &= -\frac{n}{2} \log(2\pi) + \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \right) \log(\alpha_j) \\ &\quad - \frac{1}{2} \sum_{j=1}^J \left(\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) \left(\log(v_j) + \frac{(X_i - \mu_j)^2}{v_j} \right) \right) - \lambda \left(\sum_{i=1}^n \alpha_i - 1 \right)\end{aligned}$$

Le Lagrangien admet un maximum sous la contrainte et ce maximum θ^* vérifie le système suivant :


$$\begin{cases} \frac{\partial \mathcal{L}}{\partial \alpha_j}(\theta^*) = \frac{\sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}{v_j} - \lambda & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \frac{\partial \mathcal{L}}{\partial \mu_j}(\theta^*) = \sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) (-2X_i + 2\mu_j) & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \frac{\partial \mathcal{L}}{\partial v_j}(\theta^*) = -\frac{1}{2v_j} \sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) + \frac{1}{2v_j^2} \sum_{i=1}^n \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i) (X_i - \mu_j)^2 & = 0 \quad \forall j \in \llbracket 1, J \rrbracket \\ \frac{\partial \mathcal{L}}{\partial \lambda}(\theta^*) = \sum_{i=1}^n \alpha_i - 1 & = 0 \end{cases}$$

Sous $\tilde{\theta}$ fixé, et ce qui est bien le cas, on a $\mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)$ une constante. Le système devient alors :

$$\left\{ \begin{array}{ll} \alpha_j = \frac{\sum_{i=1}^n g_{\hat{\theta}}(j|X = X_i)}{\lambda} & \forall j \in \llbracket 1, J \rrbracket \\ \mu_j = \frac{\sum_{i=1}^n X_i \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)}{\sum_{i=1}^n \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)} & \forall j \in \llbracket 1, J \rrbracket \\ v_j = \frac{\sum_{i=1}^n (X_i - \mu_j)^2 \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)}{\sum_{i=1}^n \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)} & \forall j \in \llbracket 1, J \rrbracket \\ \sum_{i=1}^J \alpha_i = 1 \end{array} \right.$$

Sous la contrainte $\sum_{j=1}^J \alpha_j = 1$ et la première équation du système précédent on obtient l'égalité suivante :

$$\begin{aligned} \sum_{j=1}^J \alpha_j &= \sum_{j=1}^J \left(\frac{\sum_{i=1}^n \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)}{\lambda} \right) \\ &= \frac{\sum_{j=1}^J \sum_{i=1}^n \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)}{\lambda} \\ &= \frac{\sum_{i=1}^n \sum_{j=1}^J \mathbb{P}_{\hat{\theta}}(Z = j|X = X_i)}{\lambda} \\ &= \frac{\sum_{i=1}^n 1}{\lambda} = 1 \end{aligned}$$

On en déduit ainsi $\lambda = n$, ainsi que le résultat énoncé. 

Tous ces inesthétiques et fastidieux calculs n'ont pas été effectué en vain. Nous les avons réalisé suite à l'introduction d'une notion nouvelle, celle de la log-vraisemblance conditionnelle ; qui elle même à été introduite faute de ne pouvoir obtenir une expression analytique de la log-vraisemblance des observations. Nous allons maintenant tâcher de mettre en exergue le rapport entre ces deux log-vraisemblances.

Chapitre 2

L'algorithme EM

Dans le présent chapitre, nous nous intéresserons A REMPLIR

2.1 Présentation laconique et pseudo-code

Dans cette partie, nous allons faire le lien entre les outils développés dans le chapitre précédent, et expliciter en quels sens ils apportent une solution à notre problématique. Rappelons que cette dernière consiste en l'estimation de paramètres d'une log-vraisemblance des observations impossible à maximiser analytiquement. Comme mentionné lors de l'introduction, nous allons présenter l'algorithme EM. Il s'agit d'une méthode itérative, constituée de deux étapes; à savoir une étape "Expectation" et une étape "Maximization". Dans la suite de ce chapitre, nous verrons que cet algorithme nous permet de calculer une suite de paramètres qui auront la particularité de faire croître la log-vraisemblance des observations à chaque itération.

2.1.1 L'étape E (Expectation)

La phase E consiste à calculer la log-vraisemblance conditionnelle $\frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{\theta}_{k-1}} [\mathcal{L}_c(\theta_{k-1}, \theta_k, X_1, \dots, X_n)]$. Son calcul sera rendu possible grâce à la formule établie dans la proposition 6 et grâce à l'utilisation des paramètres $\alpha_j, \mu_j, \sigma_j$ calculés à l'itération précédente lors de l'étape M. À l'itération zéro, il est donc indispensable de renseigner des paramètres initiaux. Leur détermination occupera donc une place centrale dans notre projet et nous en discuterons dans une section ultérieure.

2.1.2 L'étape M (Maximization)

La phase M consiste à maximiser la vraisemblance trouvée à l'étape E. Ce maximum sera atteint en $\theta_M = (\hat{\alpha}_j, \hat{\mu}_j, \hat{\sigma}_j)$ et ces trois paramètres seront calculés à l'aide maximiseur établis à la proposition 7. Une fois qu'ils auront été déterminés, ce sont ces trois paramètres qui seront utilisés dans l'itération suivante pour mettre à jour la log-vraisemblance conditionnelle lors de l'étape M.

2.1.3 La fonction simulation

Afin de pouvoir lancer l'algorithme EM, nous avons besoin d'un échantillon issue d'un mélange gaussien. Dans le cadre de ce projet, nous disposons uniquement de données répertoriant les moyennes et les écarts-types des volumes des nids de treize espèces d'oiseaux différentes. En partant de ces données (moyennes et écart-types des volumes des nids) et de la commande `rnorm(μ, σ)` de R, nous avons implémenté la fonction `simulation`, qui a pour mission de générer aléatoirement un échantillon d'un mélange Gaussien. Il s'agit d'une des plus importante fonction de ce projet, nous allons donc la décrire.

```
simulation = function(data_th, n=100){  
  X = rep(NA,n) #echantillon  
  vect_alpha = data_th[,2]  
  vect_mean = data_th[,3]  
  vect_sd = data_th[,4]  
  for(i in 1:n){
```

```

Z = runif(1)
if (Z <= vect_alpha[1]){
  X[i] = rnorm(1, vect_mean[1], vect_sd[1])
}else{
  k = 1
  l = 2
  Bool = FALSE
  cumul_alpha = cumsum(vect_alpha)
  while(Bool == FALSE){
    if((cumul_alpha[k]<=Z) & (cumul_alpha[l]>=Z)){
      Bool = TRUE
      param_index = l
    }
    k = k+1
    l = l+1
  }
  X[i] = rnorm(1, vect_mean[param_index], vect_sd[param_index])
}
}
return(X)
}

```

La fonction prend en argument *data_th* (le dataframe contenant les paramètres α , μ et σ) et n , qui indique la taille de l'échantillon à générer aléatoirement.

La fonction retourne le vecteur du mélange gaussien de taille n qui a été généré aléatoirement.

La plus partie la plus importante et subtile de ce script est celle dans laquelle nous distribuons aléatoirement les lois des différents $(X_i)_{i \in \{1, \dots, n\}}$ de l'échantillon, de sorte à avoir un bon mélange gaussien.

Pour simplifier les choses nous allons prendre un exemple dans lequel nous voulons générer un mélange de trois gaussiennes, ayant pour paramètres respectifs $\theta_1 = (\alpha_1, \mu_1, \sigma_1)$, $\theta_2 = (\alpha_2, \mu_2, \sigma_2)$ et $\theta_3 = (\alpha_3, \mu_3, \sigma_3)$. On

rappelle que $\sum_{j=1}^3 \alpha_j = 1$. Nous procédons de la manière suivante :

Nous générons une variable aléatoire Z de loi uniforme à l'aide de la commande *runif* de *R*; puis :

- Si $Z < \alpha_1$ alors $X \sim N(\mu_1, \sigma_1)$
- Sinon si $\alpha_1 \leq Z \leq \alpha_1 + \alpha_2$ alors $Z \sim N(\mu_2, \sigma_2)$
- Sinon si $\alpha_1 + \alpha_2 \leq Z \leq \alpha_1 + \alpha_2 + \alpha_3$ alors $Z \sim N(\mu_3, \sigma_3)$

Notre implémentation fonctionne dans le cas général d'un mélange de J gaussiennes avec $J \in \mathbb{N}^*$ et repose exactement sur le même principe que celui précédemment expliqué.

2.1.4 Pseudo-code de l'algorithme EM

Pour l'implémentation de cet algorithme, nous nous sommes appuyés sur le pseudo-code suivant.

Algorithm 1 L'algorithme EM (Dempster et al., 1977).

Entrée(s) : $\hat{\theta}_0 \in \Theta$, un jeu de données $x_1 \dots x_n$, $K \in \mathbb{N}$;

1: **pour** k allant de 1 à K **faire**

2: **ETAPE E :** Calculer la fonction $Q(\theta; \hat{\theta}_{k-1}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\hat{\theta}_{k-1}} [\mathcal{L}_c(\theta_k - 1, \theta_k, X_1, \dots, X_n)]$;

3: **ETAPE M :** Calculer $\hat{\theta}_k = \underset{\theta}{\operatorname{argmax}} Q(\theta; \hat{\theta}_{k-1})$;

4: **fin du pour**

5: **retourner** $\hat{\theta}_K$;

Il n'existe pas de preuve de convergence de l'algorithme EM; ce dernier peut en effet stagner dans des extremas locaux. Le choix de bons paramètres initiaux est de fait primordial. Nous verrons cela dans une prochaine section. Toutefois, nous sommes assurés que l'algorithme croît, en temoigne la section suivante.

2.2 Une preuve de la croissance

Dans cette concise partie, nous donnons une preuve de la croissance de la log-vraisemblance conditionnelle au fur et à mesure des itérations de l'algorithme EM.

Théorème 1. *Soit $(\theta_k)_{k \in \mathbb{N}}$ la suite de paramètres construite à l'aide de l'algorithme EM. La log-vraisemblance \mathcal{L}_{obs} des observations vérifie*

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \dots, X_n) \geq \mathcal{L}_{obs}(\theta_k, X_1, \dots, X_n)$$

Démonstration. Nous allons commencer cette preuve en donnant une autre forme de la log-vraisemblance, dépendant de $\mathcal{L}_{obs}(\theta, X_1, \dots, X_n)$ et d'un terme $\kappa_{\theta, \theta_k}$. Nous avons :

$$\begin{aligned} \mathcal{L}_c(\theta, \theta_k, X_1, \dots, X_n) &= \sum_{i=1}^n \sum_{j=1}^J \ln(h_\theta(X_i, j)) \mathbb{P}_{\theta_k}(Z = j | X = X_i) \\ &= \sum_{i=1}^n \sum_{j=1}^J \ln[f_\theta(X_i) \times \mathbb{P}_\theta(Z = j | X = X_i)] \mathbb{P}_{\theta_k}(Z = j | X = X_i) \\ &= \sum_{i=1}^n \sum_{j=1}^J \ln(f_\theta(X_i)) \mathbb{P}_{\theta_k}(Z = j | X = X_i) + \sum_{i=1}^n \sum_{j=1}^J \ln(\mathbb{P}_\theta(Z = j | X = X_i)) \mathbb{P}_{\theta_k}(Z = j | X = X_i) \\ &= \sum_{i=1}^n \ln(f_\theta(X_i)) \times \underbrace{\sum_{j=1}^J \mathbb{P}_{\theta_k}(Z = j | X = X_i)}_{=1} + \sum_{i=1}^n \sum_{j=1}^J \ln(\mathbb{P}_\theta(Z = j | X = X_i)) \mathbb{P}_{\theta_k}(Z = j | X = X_i) \\ &= \sum_{i=1}^n \ln(f_\theta(X_i)) + \sum_{i=1}^n \sum_{j=1}^J \ln(\mathbb{P}_\theta(Z = j | X = X_i)) \mathbb{P}_{\theta_k}(Z = j | X = X_i) \\ &= \mathcal{L}_{obs}(\theta, X_1, \dots, X_n) + \kappa_{\theta, \theta_k} \end{aligned}$$

Dès lors, on obtient

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \dots, X_n) - \mathcal{L}_{obs}(\theta_k, X_1, \dots, X_n) = \mathcal{L}_c(\theta_{k+1}, \theta_k, X_1, \dots, X_n) - \kappa_{\theta_{k+1}, \theta_k} - \mathcal{L}_c(\theta_k, \theta_k, X_1, \dots, X_n) + \kappa_{\theta_k, \theta_k}$$

Or, la quantité \mathcal{L}_c est maximisée en θ_{k+1} lors de l'étape M de l'algorithme, donc

$$\mathcal{L}_c(\theta_{k+1}, \theta_k, X_1, \dots, X_n) - \mathcal{L}_c(\theta_k, \theta_k, X_1, \dots, X_n) \geq 0$$

Il reste donc à prouver que

$$\kappa_{\theta_k, \theta_k} - \kappa_{\theta_{k+1}, \theta_k} \geq 0$$

En effet, nous avons

$$\begin{aligned}
\kappa_{\theta_k, \theta_k} - \kappa_{\theta_{k+1}, \theta_k} &= \sum_{i=1}^n \sum_{j=1}^J \ln(\mathbb{P}_{\theta_k}(Z = j|X = X_i)) \mathbb{P}_{\theta_k}(Z = j|X = X_i) \\
&\quad - \sum_{i=1}^n \sum_{j=1}^J \ln(\mathbb{P}_{\theta_{k+1}}(Z = j|X = X_i)) \mathbb{P}_{\theta_k}(Z = j|X = X_i) \\
&= \sum_{i=1}^n \sum_{j=1}^J \ln\left(\frac{\mathbb{P}_{\theta_k}(Z = j|X = X_i)}{\mathbb{P}_{\theta_{k+1}}(Z = j|X = X_i)}\right) \mathbb{P}_{\theta_k}(Z = j|X = X_i) \\
&= -n \sum_{i=1}^n \sum_{j=1}^J \ln\left(\frac{\mathbb{P}_{\theta_{k+1}}(Z = j|X = X_i)}{\mathbb{P}_{\theta_k}(Z = j|X = X_i)}\right) \mathbb{P}_{\theta_k}(Z = j|X = X_i) \times \frac{1}{n} \\
&\geq -n \times \ln\left(\sum_{i=1}^n \sum_{j=1}^J \frac{\mathbb{P}_{\theta_{k+1}}(Z = j|X = X_i)}{\mathbb{P}_{\theta_k}(Z = j|X = X_i)} \mathbb{P}_{\theta_k}(Z = j|X = X_i) \times \frac{1}{n}\right) \\
&\quad \left[\text{Cette dernière inégalité est due à la convexité de } \ln \text{ et au fait que } \sum_{i=1}^n \sum_{j=1}^J \mathbb{P}_{\theta_k}(Z = j|X = X_i) \times \frac{1}{n} = 1\right] \\
&= -n \times \ln\left(\sum_{i=1}^n \sum_{j=1}^J \mathbb{P}_{\theta_{k+1}}(Z = j|X = X_i) \times \frac{1}{n}\right) \\
&= -n \times \ln(1) \\
&= 0
\end{aligned}$$

On obtient ainsi

$$\kappa_{\theta_k, \theta_k} - \kappa_{\theta_{k+1}, \theta_k} \geq 0$$

Et finalement

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \dots, X_n) \geq \mathcal{L}_{obs}(\theta_k, X_1, \dots, X_n)$$

2

Chapitre 3

Une mise en situation

Dans ce troisième et dernier chapitre, nous nous proposons de mettre en oeuvre ce que nous avons développé dans les précédents chapitres. Nous allons nous placer en situation réelle, et étudier un jeu de données correspondant à un mélange de lois gaussiennes. Notre thématique ne changera pas, il sera toujours question de nids d'oiseaux.

3.1 Préambule

Nous n'avons hélas pu trouver des jeux de données sur la taille de nids d'oiseaux en libre accès. Toutefois, nous avons récupéré sur le site de l'*Université de Lincoln (Royaume-Uni)* l'article de recherche [4], contenant des informations fortes utiles sur une douzaine d'espèces d'oiseaux. En voici un extrait :

	Female Body Mass (g)	Total mass of nest (g)	Cup diameter parallel to long axis (mm)	Cup diameter perpendicular to long axis (mm)	Nest diameter parallel to long axis (mm)	Nest diameter perpendicular to long axis (mm)	Upper wall thickness (mm)	Base Thickness (mm)	Cup depth (mm)	Nest Height (mm)	Volume (cm ³)
Fringillidae											
European Goldfinch (<i>Carduelis Carduelis</i>) [10]	16.4	8.3 ± 2.4	62.8 ± 12.1	54.8 ± 7.4	91.4 ± 9.3	77.8 ± 7.9	12.8 ± 3.3	15.7 ± 4.3	26.0 ± 5.5	41.6 ± 7.4	38.0 ± 9.1
Common Linnet (<i>Linaria cannabina</i>) [11]	18.0	18.9 ± 5.4	74.7 ± 6.3	59.9 ± 8.6	107.9 ± 8.8	95.1 ± 10.2	16.9 ± 4.9	24.5 ± 8.9	30.6 ± 9.8	55.1 ± 9.2	60.9 ± 20.8
Common Chaffinch (<i>Fringilla coelebs</i>) [11]	21.5	14.5 ± 2.9	63.3 ± 8.1	50.8 ± 8.0	98.7 ± 10.9	90.3 ± 9.8	18.5 ± 3.6	23.6 ± 7.6	34.3 ± 7.8	58.0 ± 7.3	58.3 ± 15.0
European Greenfinch (<i>Chloris chloris</i>) [5]	25.9	22.4 ± 6.2	75.6 ± 7.8	53.9 ± 11.8	128.6 ± 13.7	99.7 ± 16.2	24.9 ± 7.9	29.4 ± 6.0	35.4 ± 5.7	64.9 ± 9.4	74.5 ± 12.2
Eurasian Bullfinch (<i>Pyrrhula pyrrhula</i>) [17]	27.3	12.1 ± 4.6	80.8 ± 12.1	66.4 ± 8.1	129.7 ± 23.4	117.5 ± 19.6	24.8 ± 10.9	24.2 ± 10.7	22.6 ± 4.5	46.8 ± 11.3	45.0 ± 3.8
Hawfinch (<i>Coccothraustes coccothraustes</i>) [4]	52.9	27.4 ± 7.3	102.2 ± 17.9	78.8 ± 25.2	153.4 ± 19.1	131.3 ± 27.1	25.4 ± 5.9	23.3 ± 4.9	31.4 ± 10.9	54.7 ± 11.5	71.6 ± 12.9

Parmi elles ce trouve des volumes moyens de nids ainsi que la variance associée.

Afin de se conformer aux hypothèses que nous avons précédemment mentionné, nous supposerons ici que les volumes des nids suivent une loi normale. Nous ferons une dernière hypothèse ; nous supposerons le nombre de lois est connu lors de l'observation.

Plaçons nous dès à présent dans le cadre d'une étude. Pour ce faire, commençons par établir une méthodologie. Il sera en premier lieu choisi, aléatoirement, plusieurs espèces d'oiseaux parmi celles que nous avons à disposition. Nous allons générer, avec R , les données dont nous avons besoin. Cet n -échantillon sera créé grâce à notre fonction de simulation [A DECRIRE ? ?]. Notre jeu de données étant prêt, nous pourrons commencer l'étude.

Le premier problème qui se posera sera celui du choix des paramètres initiaux, notre algorithme nécessitant des valeurs initiales. Nous avons diverses solutions à notre portée. La première est de faire une exploration des données ; via une représentation graphique de la densité de l'échantillon. Si des pics bien distincts se présentent,

cela nous permettra de régler judicieusement les paramètres initiaux. Une autre solution est d'utiliser notre fonction de détermination des paramètres, qui comme nous l'avons vu au précédent chapitre est basée sur ????. Une fois cela fait, nous pouvons exécuter notre algorithme et obtenir son estimation des paramètres. Nous pourrions dès lors émettre les conclusions appropriées. Nous partons, pour ainsi dire, à l'aveugle, mais il sera gardé en mémoire les vraies valeurs, afin de vérifier si nous avons abouti aux bonnes conclusions.

Nous avons, à partir du tableau de données sus-mentionné, construit le data-frame *nest_data* reprenant les noms des espèces, ainsi que les volumes moyens et variances associées. Ce dernier nous servira à créer les échantillons

3.2 Étude de simulations

Les différentes sections ci-dessous auront vocations à présenter plusieurs simulations dans le but de tester l'efficacité de notre implémentation de l'algorithme EM.

3.2.1 Étude d'un mélange gaussien à forte séparations

Dans cette partie, nous nous intéresserons à un mélange de trois lois gaussiennes, ayant comme lois respectives $\mathcal{N}(\mu_1, \sigma_1)$, $\mathcal{N}(\mu_2, \sigma_2)$ et $\mathcal{N}(\mu_3, \sigma_3)$ bien séparées, c'est à dire avec des paramètres bien différents les uns des autres (i.e. $|\alpha_i - \alpha_j|_{i \neq j}$; $|\mu_i - \mu_j|_{i \neq j}$; $|\sigma_i - \sigma_j|_{i \neq j}$ assez grands). Nous avons généré un échantillon d'un mélange de trois gaussiennes à fortes séparations de taille $n = 500$. Les trois lois du mélange sont les suivantes :

- $\mathcal{N}(11, 1)$
- $\mathcal{N}(49, 10)$
- $\mathcal{N}(92, 3)$

La figure ci-dessous représente la distribution du mélange

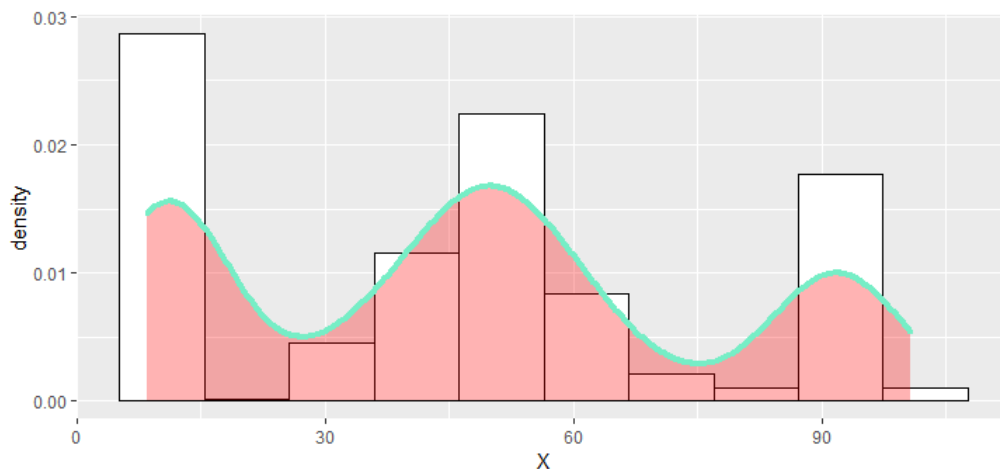


FIGURE 3.1 – Mélange gaussien à forte séparation

Nous avons lancé l'algorithme EM en choisissant un nombre d'itération $k = 10$. Comme on peut le voir sur la capture d'écran ci-dessous, les paramètres estimés par l'algorithme EM sont assez proche de ceux théoriques.

```
> print(good_dfTestTh3)
  bird_names proportion_alpha mean_th sd_th
1 Good species 1          0.3      11     1
2 Good species 2          0.5      49    10
3 Good species 3          0.2      92     3
> algo_EM(good_dfTestInit3, xGood3, 10)
  bird_names  alpha      mu  sigma
1 Good species 1 0.2939353 11.05072 0.9611203
2 Good species 2 0.5042109 49.33787 9.7178553
3 Good species 3 0.2018537 91.92101 2.9731759
```

FIGURE 3.2 – Paramètres théoriques VS paramètres estimés

Nous allons maintenant regarder si l'augmentation de la taille de l'échantillon du mélange gaussien généré aléatoirement aura pour effet de diminuer et l'erreur absolue entre les paramètres estimés par l'algorithme EM, et les valeurs théoriques dans le cas d'un mélange à forte séparation. Pour cela nous allons utiliser notre fonction *Monte-Carlo*.

Nous avons effectué 100 itérations de Monte-Carlo pour les taille d'échantillon suivante :

— $n = 100$

— $n = 250$

— $n = 500$

Les boîtes à moustaches (boxplot) ci-dessous, modélisent l'erreur absolue entre les paramètres estimés par l'algorithme Em et ceux théoriques.

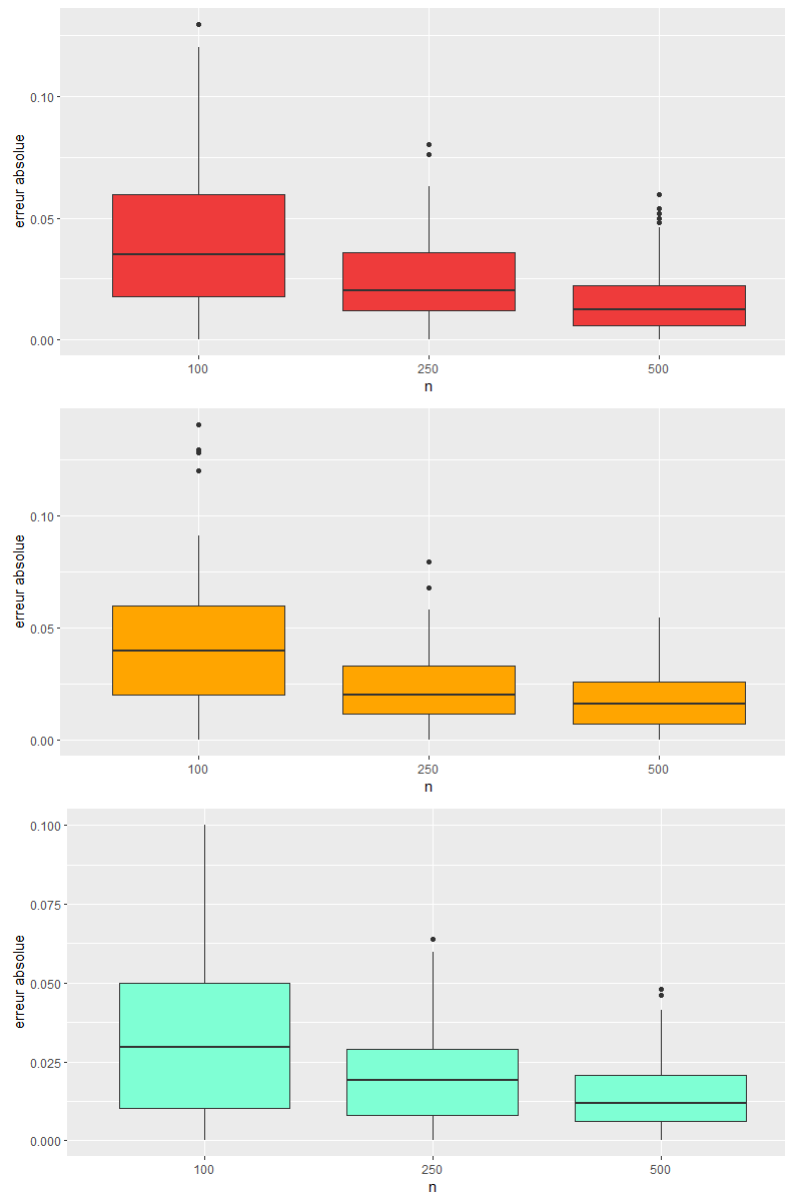


FIGURE 3.3 – Boxplot des erreurs absolues pour α_1 , α_2 et α_3

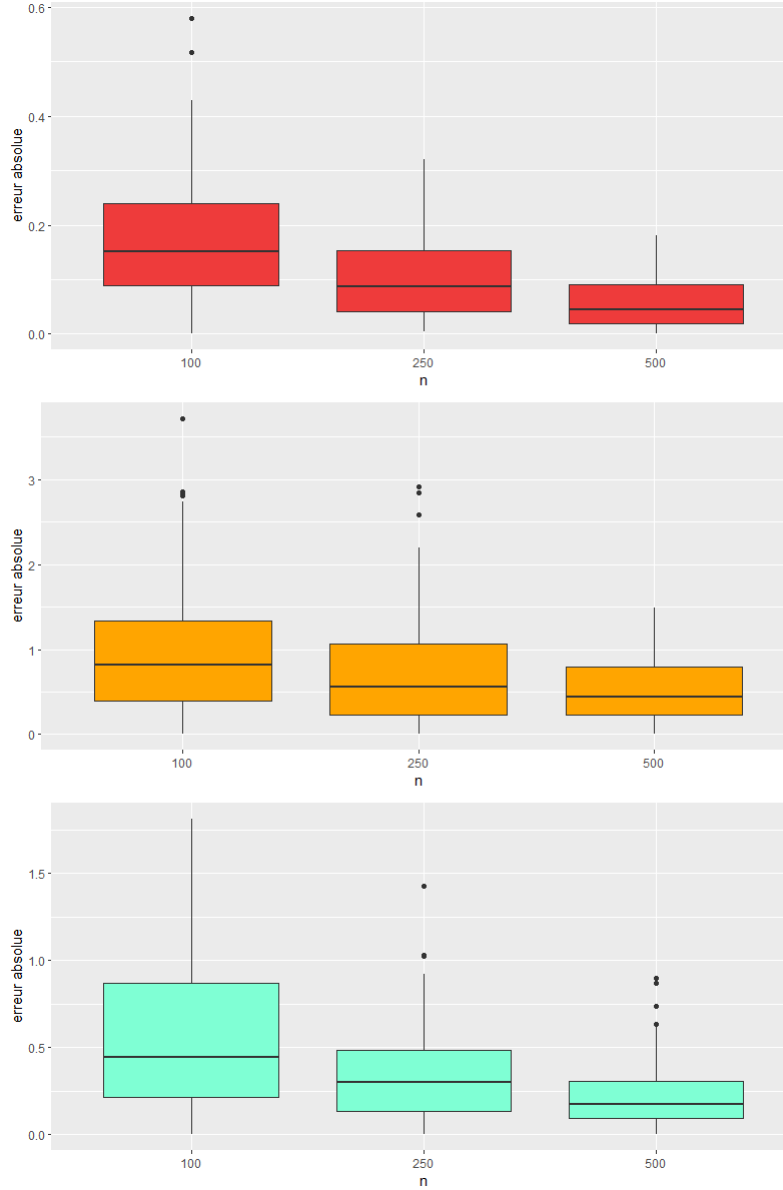


FIGURE 3.4 – Boxplot des erreurs absolues pour μ_1 , μ_2 et μ_3

D'après les boites à moustaches, nous constatons que plus la taille de l'échantillon du mélange gaussien est grande, et plus l'erreur absolue entre les paramètres estimés par l'algorithme EM et ceux théoriques, diminuent.

3.2.2 Étude d'un mélange gaussien à faible séparations

Dans cette partie, nous nous intéresserons à un mélange de trois gaussiennes ayant comme loi respectives $\mathcal{N}(\mu_1, \sigma_1)$, $\mathcal{N}(\mu_2, \sigma_2)$ et $\mathcal{N}(\mu_3, \sigma_3)$ faiblement séparées, c'est à dire avec des paramètres assez proches les uns des autres (i.e. $|\alpha_i - \alpha_j|_{i \neq j}$; $|\mu_i - \mu_j|_{i \neq j}$; $|\sigma_i - \sigma_j|_{i \neq j}$ assez petits). De même que pour la première étude, nous avons généré un échantillon d'un mélange de trois lois gaussiennes de taille 500, cette fois-ci avec faible séparation. Les trois lois du mélange sont les suivantes :

- $\mathcal{N}(5, 1)$
- $\mathcal{N}(13.6, 11)$
- $\mathcal{N}(15.3, 6)$

La figure ci-dessous représente la distribution du mélange à faibles séparations ;

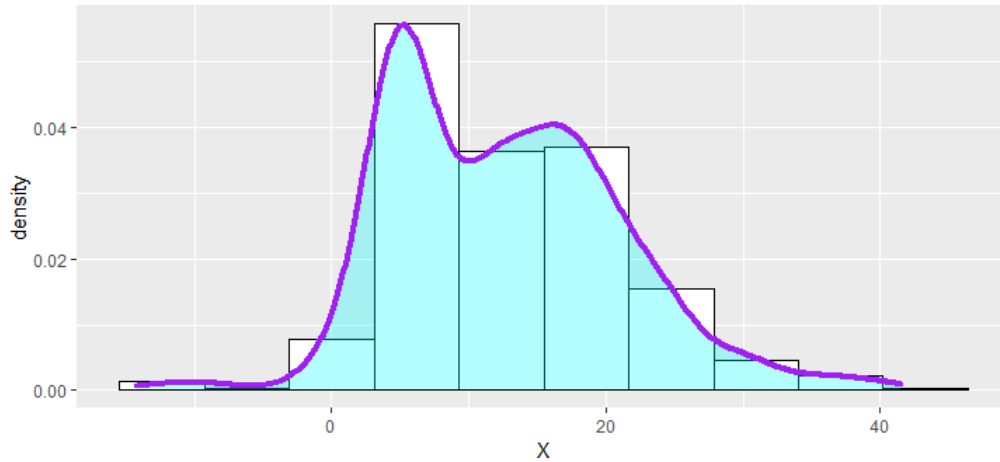


FIGURE 3.5 – Mélange gaussien à faibles séparations

Ici aussi, nous avons lancé l'algorithme EM en choisissant un nombre d'itération $k = 10$. Comme nous pouvons le voir sur la capture d'écran ci-dessous, même avec un mélange à faibles séparations, les paramètres estimés par l'algorithme EM sont assez proches de ceux théorique.

```
> print(bad_dfTestTh3)
  bird_names proportion_alpha mean_th sd_th
1 Bad species 1      0.24      5.0    1
2 Bad species 2      0.36     13.6   11
3 Bad species 3      0.40     15.3    6
> algo_EM(bad_dfTestInit3, xBad3, 10)
  bird_names   alpha      mu   sigma
1 Bad species 1 0.2508061  4.93471 1.143349
2 Bad species 2 0.5426620 14.56490 8.936383
3 Bad species 3 0.2065318 16.73281 5.410275
```

FIGURE 3.6 – Paramètres théoriques VS paramètres estimés

De même que pour l'étude précédente, nous allons maintenant regarder si l'augmentation de la taille de l'échantillon du mélange gaussien généré aléatoirement aura pour effet de diminuer l'erreur absolue entre les paramètres estimés par l'algorithme EM et les valeurs théoriques dans le cas d'un mélange à faibles séparations. Pour cela nous allons utiliser notre fonction *Monte-Carlo*.

Nous avons effectué 100 itérations de Monte-Carlo pour les différentes tailles d'échantillons suivantes :

— $n = 100$

— $n = 250$

— $n = 500$

Les boîtes à moustaches (boxplot) ci-dessous, modélisent l'erreur absolue entre les paramètres estimés par l'algorithme Em et ceux théoriques dans le cas d'une faible séparation.

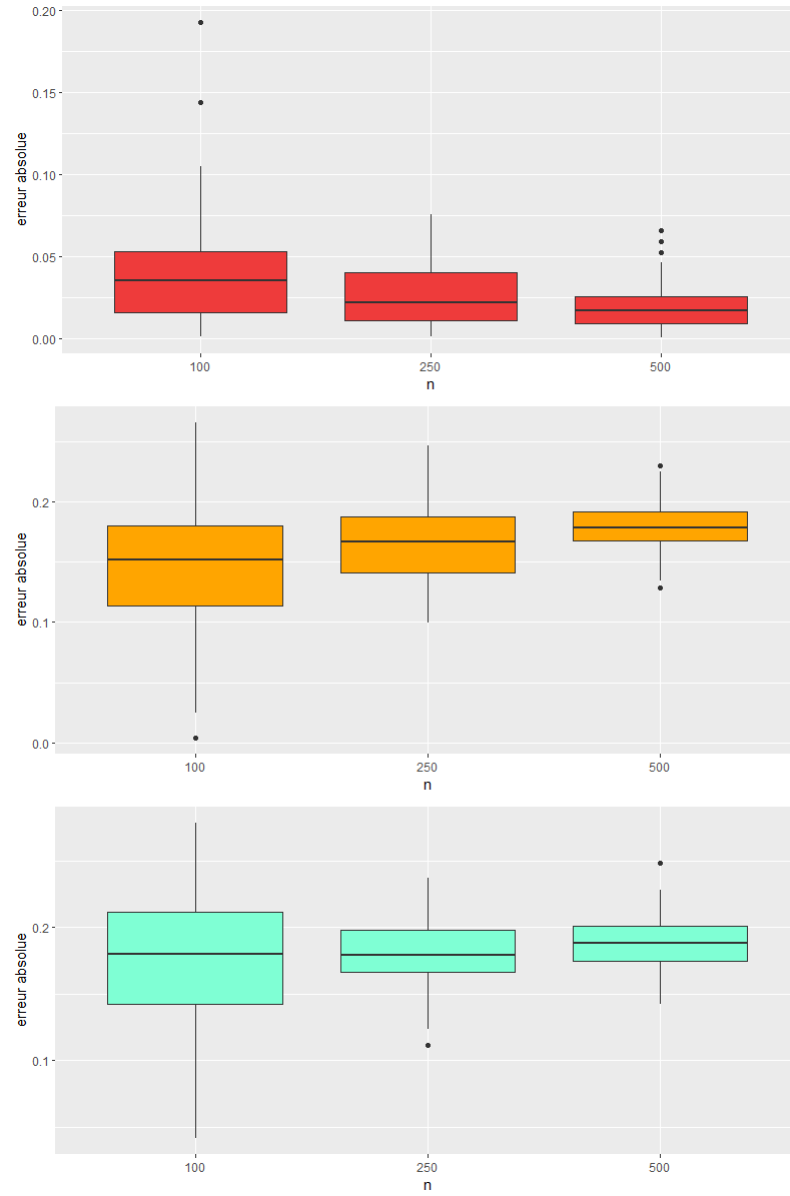


FIGURE 3.7 – Boxplot des erreurs absolues pour α_1 , α_2 et α_3

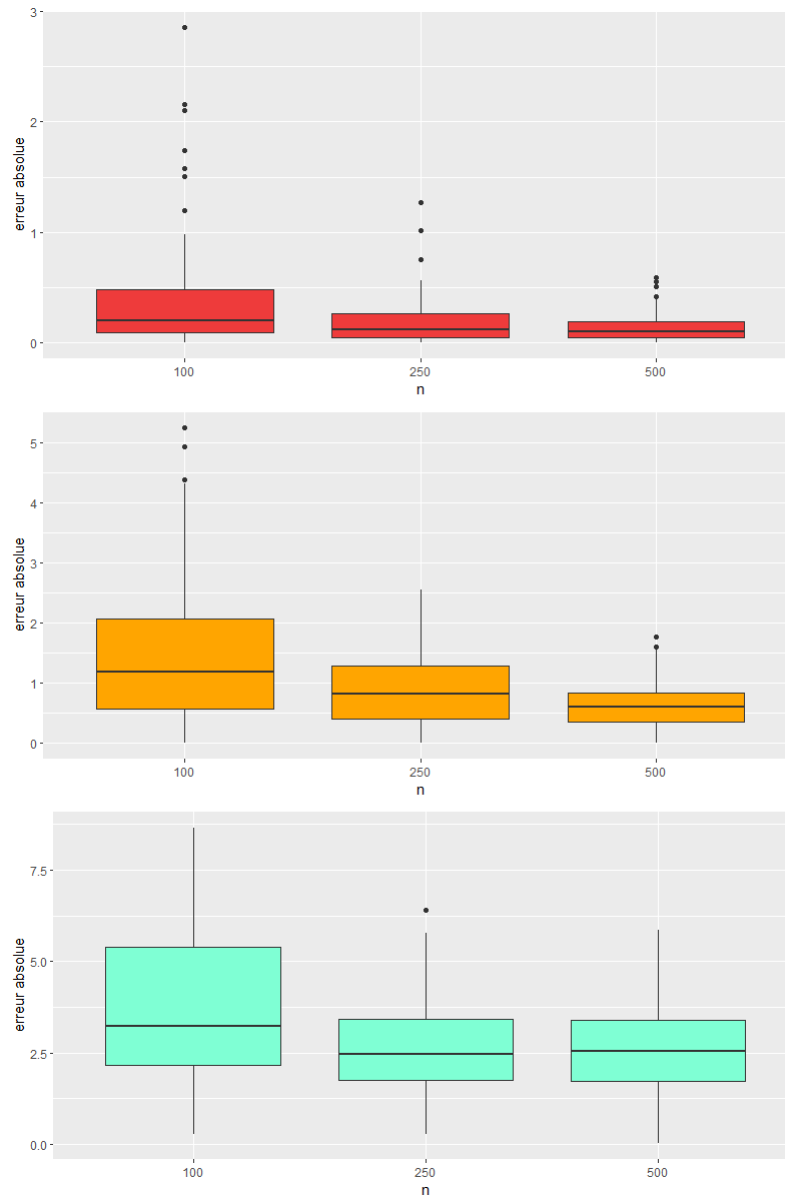


FIGURE 3.8 – Boxplot des erreurs absolues pour μ_1 , μ_2 et μ_3

Même dans le cas d'un mélange à variables faiblement séparées, on constate que plus on augmente la taille de l'échantillons et plus l'erreur absolue diminue, ce qui est rassurant.

3.3 Étude de cas avec les vrais données

3.3.1 Cas d'un mélange à deux lois avec des vrais données

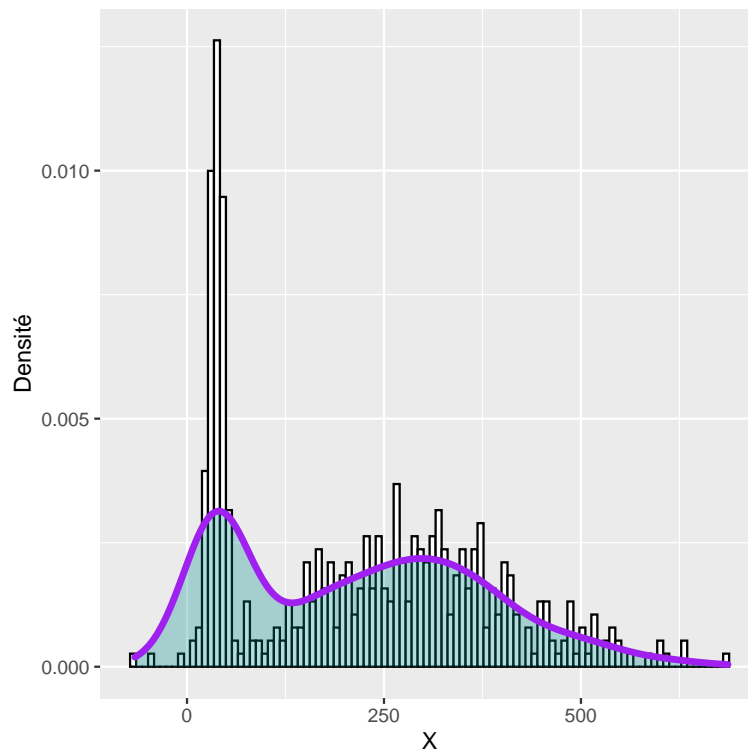
Dans cette section, nous allons étudier le cas d'un mélange de deux lois. Commençons par sélectionner aléatoirement deux espèces, ainsi que les proportions associées. Pour ce faire, nous utiliserons notre fonction, la fonction `random_species`.

```
df <- random_species(nest_data, 2)
```

Nous pouvons dès lors générer l'échantillon, à l'aide de notre fonction `simulation`.

```
data <- simulation(df, 500)
```

Le premier réflexe qui se présente est de tracer la fonction de densité associée à l'échantillon :



Le mélange des deux lois se distingue au premier coup d'oeil ; nous distinguons nettement deux pics. L'"étalement" des pics n'étant pas les mêmes, nous pouvons supposer que les variances des deux lois sont différentes. Il est cependant difficile d'émettre des hypothèses quant aux proportions.

Commençons l'étude rigoureuse.

Nous devons pour commencer déterminer les paramètres initiaux. Nous allons utiliser diverses méthodes suivant les paramètres.

Pour les deux moyennes initiales, nous allons utiliser les quantiles ; et plus précisément le premier et le troisième. Travaillant sur R , nous pouvons aisément obtenir ces derniers :

```
quantile(data)
```

0%	25%	50%	75%	100%
-89.38908	44.07317	236.24277	344.55731	736.45904

Nous obtenons ainsi nos moyennes initiales : $\mu_{1_{init}} = 44.07317$ et $\mu_{2_{init}} = 344.55731$.

Pour l'initialisation des écarts-types, nous prendrons $\sigma_{1_{init}} = \sigma_{2_{init}} = \hat{\sigma}$; où

$$\hat{\sigma} := \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

est l'écart-type empirique. Une fois de plus, R nous permet de calculer cela aisément, grâce à la fonction `sqrt(var(data))`. Nous obtenons

```
[1] 164.8838
```

Ainsi, $\sigma_{1_{init}} = \sigma_{2_{init}} = 164.8838$.

Pour les proportions initiales, nous proposons de nouveau de les prendre égales, ainsi $\alpha_{1_{init}} = \alpha_{2_{init}} = 0.5$. On construit le dataframe des paramètres initiaux :


```
param_init = data.frame(bird_names = c("European Goldfinch", "Ring Ouzel"),
                        alpha_init = c(0.5, 0.5), mean_init = c( 44.07317,
                        344.55731),
                        sd_init = c(164.8838, 164.8838))
```

Enfin, nous pouvons lancer notre implémentation de l'algorithme EM. Comme nous l'avons vu dans le précédent chapitre, une dizaine d'itérations suffisent pour obtenir de bons résultats.

```
algo_EM(param_init, data, 10)
```

	bird_names	alpha	mu	sigma
1	European Goldfinch	0.3322016	38.50204	16.76789
2	Ring Ouzel	0.6677984	312.82254	124.46222

Nous obtenons ainsi nos paramètres estimés ; comparons le avec les vrais valeurs. Ces dernières sont contenues dans le dataframe *df*.

	bird_names2	proportion_alpha	mean_volume	sd_volume
1	European Goldfinch	0.2878713	38.0	9.1
12	Ring Ouzel	0.7121287	298.6	125.1

Les proportions sont toutes deux très bien estimées, les erreurs sont de l'ordre de 5%, ce qui est formidable. La moyenne de la première espèce est presque égale à la moyenne théorique, l'erreur est inférieure à 0.5%. Celle de la seconde est un peu moins bien estimée, mais l'erreur d'estimation est de l'ordre de 3%, ce qui est très satisfaisant. Concernant les variances, la première présente une erreur d'estimation de l'ordre de 7%, ce qui est plutôt bon. La seconde variance est elle très bien estimée, avec une erreur presque nulle.

Pour ce premier exemple de mélange de lois, nous obtenons de très bonne approximations.

3.4 Cas d'un mélange à trois lois avec des vrais données

Dans cette section, nous étudierons le cas d'un mélange à trois lois.

Bibliographie

- [1] Doc de Brunel???
- [2] <https://members.loria.fr/moberger/Enseignement/AVR/Exposes/algo-em.pdf>
- [3] http://faculty.washington.edu/fxia/courses/LING572/EM_collins97.pdf
- [4] <https://core.ac.uk/download/pdf/155777956.pdf>
- [5] <https://cran.r-project.org/web/packages/mclust/mclust.pdf>
- [6] https://rstudio-pubs-static.s3.amazonaws.com/154174_78c021bc71ab42f8add0b2966938a3b8.html

Annexes

Annexe A

Le package *mclust*

Nous avons, dans un élan d'audace, commencé par programmer à la main l'algorithme EM, en nous appuyant sur le pseudo-code explicité en première partie du chapitre II.

Cependant, il existe une librairie *R* - la librairie *mclust* - contenant une implémentation de l'algorithme EM. Notre algorithme étant fonctionnel, nous ne détaillerons pas ici le fonctionnement de ce Package. Il est néanmoins pertinent de l'expérimenter, voire de comparer ces résultats avec ceux notre algorithme. Nous reprendrons ici les espèces étudiées lors du dernier chapitre ; les divers paramètres seront conservés, seul l'échantillon généré changera. Nous nous sommes appuyés sur [5] et [6] afin d'obtenir les éléments nécessaires à l'utilisation de ce package.

Pour commencer, installons et chargeons le Package *mclust*.

```
install.packages("mclust")
library("mclust")
```

Nous reprenons les données des nids d'oiseaux :

```
bird_names = c("European Goldfinch", "Common Linnet", "Common Chaffinch",
               "European Greenfinch", "Eurasian Bullfinch", "Hawfinch",
               "Stonechat", "European Robin", "Whinchat", "Song Thrush",
               "Common Blackbird", "Ring Ouzel", "Mistle Thrush")

mean_volume = c(38.0, 60.9, 58.3, 74.5, 45.0, 71.6, 91.0, 68.4, 51.9, 288.9,
               293.6, 298.6, 266.1)

sd_volume = c(9.1, 20.8, 15.0, 12.2, 3.8, 12.9, 46.5, 29.8, 27.4, 55.9,
              78.5, 125.1, 56.6)
```

Puis, il suffit de construire des dataframe. Ici, nous considérerons deux mélanges ; un mélange à deux lois et un autre à trois lois.

```
df_2= data.frame(bird_names = c("European Goldfinch", "Ring Ouzel"),
                 ,proportion_alpha = c(0.3, 0.7), mean = c(38, 298.6),
                 sd = c(9.1, 125.1))

df_3= data.frame(bird_names = c("Common Linnet", "Common Chaffinch", "Hawfinch" )
                 ,proportion_alpha = c(0.6, 0.3, 0.1), mean = c(60.9, 58.3,
                 71.6),
                 sd = c(20.8, 15.0, 12.9))
```

Nous reprenons ici notre propre fonction de simulation

```
simulation = function(data_th, n=100)
```

Les prérequis étant posés, nous simulons un échantillon $X2$ de deux espèces d'oiseaux et un autre $X3$ de trois espèces d'oiseaux :

```
set.seed(1907)
X2 <- simulation(df_2)
X3 <- simulation(df_3)
```

Le Package *mclust* est des plus complet ; les possibilités étant très vastes et hors du cadre de ce projet (notamment les fonctionnalités de clustering), nous regarderons uniquement la fonction qui nous intéresse, à savoir la fonction *densityMclust*.

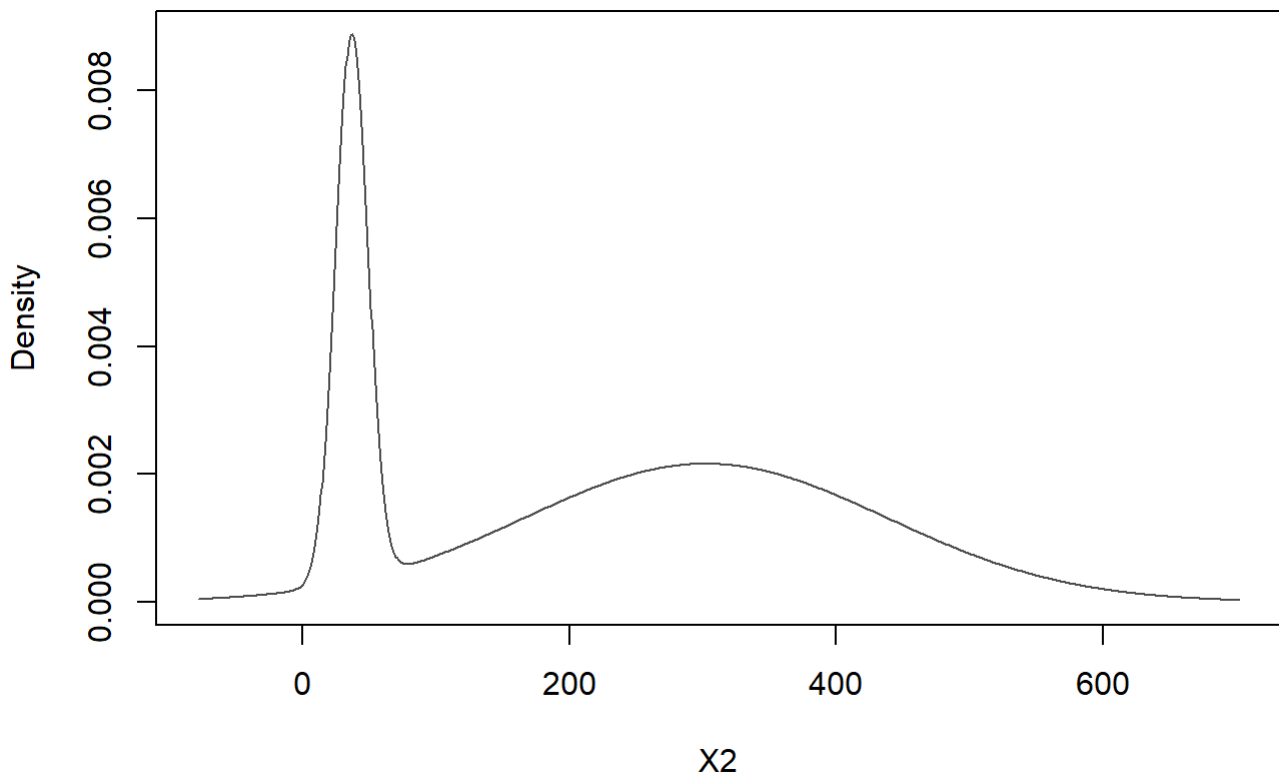
Cette dernière prend en argument des fonctionnalités pertinentes, comme le nombre de mélanges, mais ne permet pas de régler manuellement des valeurs initiales pour les paramètres à estimer.

A.1 Un exemple sur un mélange à deux lois

Commençons par un exécuter la fonction *densityMclust* sur notre exemple de mélange à deux lois, contenu dans le dataframe $X2$:

```
est_2 <- densityMclust(X2)
```

Il est en premier lieu retourné le graphe de la densité du mélange de lois.



Nous pouvons nettement distinguer les deux "pics", correspondant aux deux gaussiennes mélangées. Intéressons-nous maintenant à l'objet créé *est_2*.

```
est_2
```

```
'densityMclust' model object: (V,2)
```

Available components:

[1]	"call"	"data"	"modelName"	"n"	"d"
[6]	"G"	"BIC"	"loglik"	"df"	"bic"
[11]	"ic1"	"hypvol"	"parameters"	"z"	"classification"
[16]	"uncertainty"	"density"			

Ici nous voulons les paramètres estimés, nous nous concentrerons donc que sur la treizième coordonnée de ce vecteur.

Rappelons que les divers paramètres de ce mélange sont : 0.3 et 0.7 en proportions; 38 et 298.6 pour les moyennes; et 9.1 et 125.1 en écart-types.

```
print("Proportions estimées:")
est_2[13]$parameters$pro
print("Moyennes estimées:")
est_2[13]$parameters$mean
print("Ecart-types estimés:")
(est_2[13]$parameters$variance$sigmaSq)^(1/2)
```

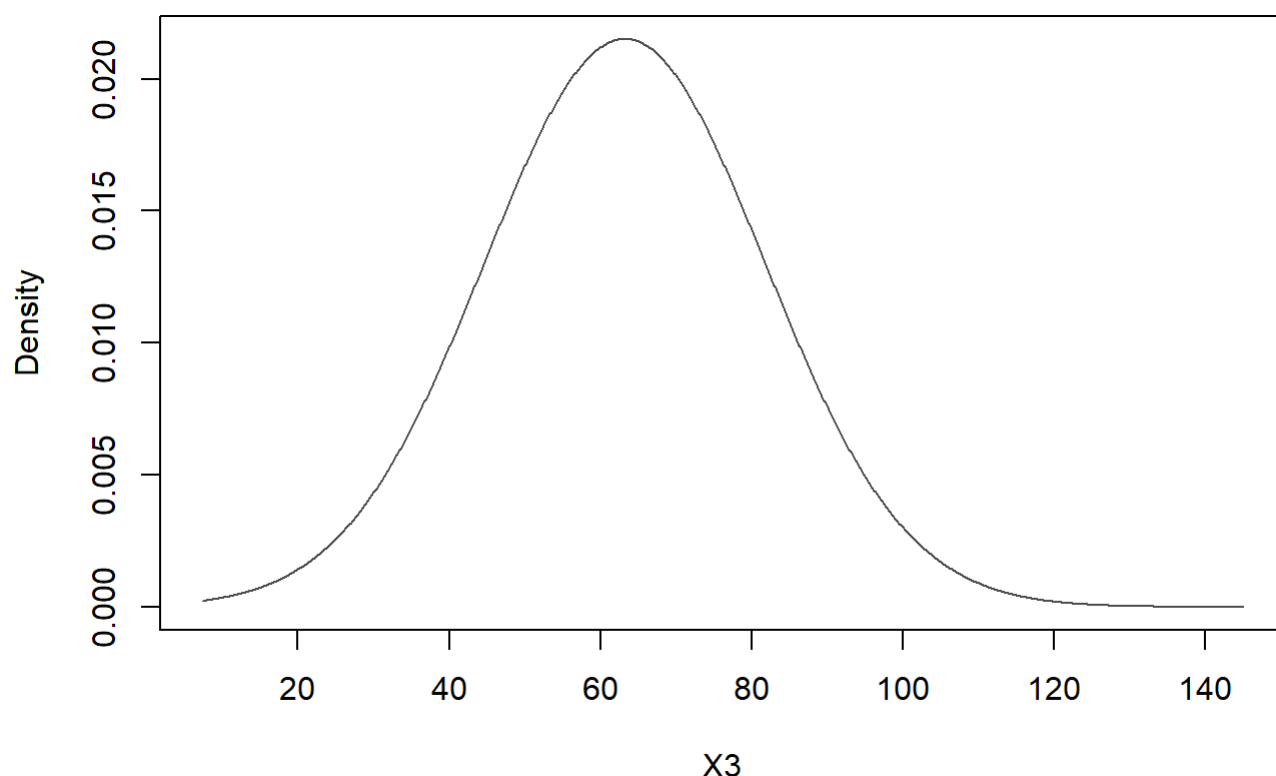
```
[1] "Proportions estimées:"
[1] 0.2612243 0.7387757
[1] "Moyennes estimées:"
   1      2
36.80898 301.97108
[1] "Ecart-types estimés:"
[1] 12.16256 136.32256
```

Ici, le nombre de mélange est exact. Les proportions sont très bien estimées, l'erreur la plus importante est de l'ordre de 12%. Il en va de même pour les moyennes, les erreurs sont d'ordres inférieures à 10%. Les variances sont elles aussi très bien estimées, les erreurs sont négligeables.

A.2 Un exemple sur un mélange à trois lois

Regardons maintenant le cas d'un mélange de trois lois. Afin de mettre à rude épreuve l'algorithme, nous allons choisir les espèces telles que les moyennes et variances soient proches. Les proportions seront quant à elles bien distinctes, nous allons voir pourquoi.

```
est_3 <- densityMclust(X3)
```



Nous obtenons ici quelque chose d'intéressant ; la fonction de densité de ce mélange de trois lois paraît toute à fait gaussienne. Sans une exploration plus approfondie des données, nous commettrions une chagrinante erreur et des conclusions totalement faussées...

Il est ici pertinent d'observer la structure des données ; plus précisément, nous allons effectuer un test de *Shapiro*.

```
shapiro.test(X3)
```

Shapiro-Wilk normality test

data: X3

W = 0.97982, p-value = 0.1287

La p-value est de 0.1287, ce qui est certes peu élevée, mais pas assez pour rejeter l'hypothèse (H_0) de normalité. Nous sommes ici dans une situation ambiguë.

Observons maintenant comment *densityMclust* se défend face à cette situation.

Rappelons que les divers paramètres de ce mélange sont : 0.6, 0.3 et 0.1 en proportions ; 60.9, 58.3 et 71.6 en moyenne ; et 20.8, 15 et 12.9 en écart-types.

```
print("Proportions estimées:")
est_3[13]$parameters$pro
print("Moyennes estimées:")
```

```
est_3[13]$parameters$mean
print("Ecart-types estimés:")
(est_3[13]$parameters$variance$sigma^2)^(1/2)
```

```
[1] "Proportions estimées:"
[1] 1
[1] "Moyennes estimées:"
[1] 63.20547
[1] "Ecart-types estimés:"
[1] 18.54033
```

Le premier élément notable est que l'algorithme échoue à établir le nombre correct de lois. L'unique moyenne et écart-type estimés ne sont quant à eux pas absurde.

Nous allons relancer la fonction sur le même jeu de données, en précisant cette fois-ci le nombre de lois.

```
est_3b <- densityMclust(X3, G = 3)
print("Proportions estimées:")
est_3b[13]$parameters$pro
print("Moyennes estimées:")
est_3b[13]$parameters$mean
print("Ecart-types estimés:")
(est_3b[13]$parameters$variance$sigma^2)^(1/2)
```

```
0.6, 0.3, 0.1 60.9, 58.3, 71.6 20.8, 15, 12.9
```

```
[1] "Proportions estimées:"
[1] 0.2552686 0.5642272 0.1805042
[1] "Moyennes estimées:"
      1      2      3
56.63179 62.36268 75.13635
[1] "Ecart-types estimés:"
[1] 17.51051
```

Les proportions sont plutôt bien estimées, quoique légèrement surestimées pour deux d'entre elles, mais les erreurs restent faibles. Il en est étonnant de même pour les moyennes, qui sont très bien estimées. Ceci est surprenant au vu de l'allure de la densité. Cependant, il n'est estimé qu'un unique écart-type, ce qui n'est guère étonnant. Notons que celle-ci est à peu près égale à la moyenne des écart-types des différentes lois.

Ce cas ambigu met en exergue les limites de l'algorithme implémenté dans ce package.